From the DIVISION OF MOLECULAR NEUROBIOLOGY
DEPARTMENT OF MEDICAL BIOCHEMISTRY AND BIOPHYSICS
Karolinska Institutet, Stockholm, Sweden

Counting molecules in cell-free DNA and single cells RNA

Kasper Karlsson

Karolinska Institutet

Stockholm 2016

The cover painting is designed by Dongjiao Karlsson

"If you can control your passions,
it's due to your passions weaknesses,
not your strength".

–Jeanne d'Arc

Published by Karolinska Institutet. **Printed by E-Print AB 2016**

 ISBN 978-91-7676-220-2

ABSTRACT

The field of Molecular Biology got started in earnest with the discovery of the molecular structure of DNA. This lead to a surge of interest into the relationships between DNA, RNA and proteins, and to the development of fundamental tools for manipulating those substances, such as cutting, ligating, amplifying, visualizing and size-selecting DNA. With these tools at hand it was possible to begin sequencing DNA, a process that took a leap forward in 2005 with the advent of Next Generation Sequencing (NGS). An inherent problem with NGS is that both the sequencing process and the library preparation introduce errors and biases. The massive amount of data generated by NGS, and the use of NGS in clinical settings, has created a demand for methods that can account for this and thereby making the sequencing data correct and reproducible.

Part 1 of this thesis briefly describes the development of Molecular Biology from the discovery of the molecular structure of DNA until today. Part 2 describes the development of error correcting and molecule counting methods, and part 3 describes the results of the papers of the thesis.

Paper I introduces the concept of Unique Molecular Identifiers (UMI) for reducing noise in molecular karyotyping and RNA sequencing data. Paper II compares the use of UMIs, a novel amplification-free method, and standard library preparation in Non-Invasive Prenatal Testing (NIPT) of fetal karyotype. Paper III uses the UMI concept together with single-cell tagged reverse transcription (STRT) to examine promoter preference in single cells. Finally paper IV use UMIs combined with PacBio sequencing to examine the full-length transcriptome in single cells.

KEY WORDS

UMI, amplification-free library preparation, NIPT, alternative promoter usage, isoforms, single-cell, STRT

LIST OF PUBLICATIONS

I. Kivioja, T., Vaharautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2012).
Counting absolute numbers of molecules using unique molecular identifiers.
Nat Methods *9*, 72-74.

II. Karlsson, K., Sahlin, E., Iwarsson, E., Westgren, M., Nordenskjöld, M., and Linnarsson, S. (2015).
Amplification-free sequencing of cell-free DNA for prenatal non-invasive diagnosis of chromosomal aberrations.
Genomics *105*, 150-158.

III. Kasper Karlsson, Peter Lönnerberg, Sten Linnarsson
Alternative Promoters are co-regulated in single cells
Manuscript

IV. Kasper Karlsson, Sten Linnarsson
Single-cell mRNA isoform diversity in the mouse brain
Manuscript

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

A	Adenine
bp	Base Pair
C	Cytosine
CA1	Cornu Ammonis area 1
CAGE	Cap Analysis of Gene Expression
CAS	Consensus sequences plus Adjusted Singletons
cfDNA	cell-free DNA
ChR2	Channelrhodopsin 2
CV	Coefficient of Variation
dATP	Deoxyadenosine triphosphate
DBR	Degenerate Base Regions
DCS	duplex consensus sequences
ddN	dideoxynucleotide
DMS	Different Molecular Species
DNA	Deoxyribonucleic acid
dTTP	Deoxythymidine triphosphate
ENCODE	Encyclopedia of DNA Elements
ERCC	External RNA Controls Consortium
FANTOM	Functional annotation of the mammalian genome
G	Guanine
GC	Guanine-Cytosine
HPG	Human Genome Project
iCLIP	UV cross-linking and immunoprecipitation
LEA-Seq	Low-Error Amplicon Sequencing
Mbp	Mega base pair
MIGEC	Molecular Identifier Groups-based Error Correction
mRNA	messenger RNA
NGS	Next Generation Sequencing
NIPT	Non-Invasive Prenatal Testing
PacBio	Pacific Biosciences
PCR	Polymerase Chain Reaction
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
SBS	Sequencing By Synthesis
Safe-Seqs	Safe-Sequencing System
SMRT	Single Molecule Real Time
STRT	Single-cell Tagged Reverse Transcription
T	Thymine
TSS	Transcription Start Site
TCGA	The Cancer Genome Atlas
UID	Unique Identifier
UMI	Unique Molecular Identifiers

1 MOLECULAR BIOLOGY

The blueprint of human life is found in the nucleus of each cell in our body and consists of 23 chromosome pairs containing two deoxyribonucleic acid (DNA) strands. The DNA molecule is built from sugar, phosphate and four variable bases: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). The sugar and phosphate make up the backbone of the DNA molecule, while the bases contain the genetic information. The combination of a sugar, phosphate and a base is the building block of DNA and is referred to as a nucleotide. The two strands of DNA are complementary, the base A pairs with T, and C pairs with G forming what is called a base pair (bp). The structure of DNA makes it suitable for copying as Watson and Crick stated in their seminal paper: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material" (1).

The DNA code is written with the four nucleotide letters A, C, G and T, and a typical gene is around 10-15 kilobases (kb) long. While DNA contains information, it cannot perform any chemical work. The main functional units in cells are the proteins. To produce proteins, DNA is transcribed into a similar molecule, ribonucleic acid (RNA). The RNA molecule is then subjected to post-transcriptional modifications that turns it into a mature RNA molecule, which can be substantially shorter than the pre-mRNA. The RNAs that produce proteins are called messenger RNA (mRNA) and the average size of an mRNA molecule is around 2 kb. The transformation of mRNA to proteins is performed by the ribosome, and is referred to as translation. Proteins are translated by reading three consecutive RNA bases (codons) encoding one amino acid, and coupled serially to form a chain of amino acids. The amino acid chain folds into a three-dimensional structure, and is subjected to post-translational modifications, yielding a protein ready to perform its designated function.

The central dogma of molecular biology, postulated by Francis Crick in 1958, states that information flows in the direction from DNA to RNA to proteins. It also states that DNA can copy itself and that in special cases RNA can copy itself and make DNA, as has been seen in some viruses. However once a protein is formed it cannot be transformed back to RNA or DNA, and neither can it copy itself (2, 3). The Central Dogma illustrates the key difference between molecular biology and the science from which it emerged, biochemistry. Whereas the latter is concerned with the flow of matter and energy, molecular biology is primarily concerned with the flow of information.

The DNA in all cells of an individual is identical with the exception of rare mutations that occur mainly during replication, so-called somatic mutations. Still the phenotype of distinct cell types can be very different. This apparent contradiction is possible because specific regulatory programs are active in different cells and control what genes are expressed. No cell expresses all genes, and depending on which genes are expressed the cells will have different phenotypes and be able to perform different functions.

After it was discovered that DNA was the substance that carried the heritability in cells, substantial effort was made to be able to read and manipulate the DNA code. Many new techniques were developed and some of them are still used today. This fascinating period and the discoveries made has been described in numerous books and nicely summarized in a review by Supratim Choudhuri (4).

1.1 TOOLS USED IN MOLECULAR BIOLOGY

To be able to separate and identify nucleic acids, gel electrophoresis was developed (5). It takes advantage of the negative charge of nucleic acids to drive them through a polyacrylamide or agarose gel by applying a current. Smaller molecules travel more quickly through the gel and can be separated from larger molecules. Molecules in the gel can also be visualized e.g. by dyes and ultraviolet light. Edwin Southern developed an application of this by transferring the gel-separated DNA fragments to a paper and then hybridizing known radioactive RNA probes to the fragments allowing detection of specific DNA sequences. This became known as southern blotting (6). A similar technique was later developed for RNA and called Northern blotting (7). Southern blotting and many other methods depended on a technique that could cut DNA into smaller pieces. In 1970 Hamilton Smith and colleagues isolated a so called restriction enzyme from *Haemophilus influenza (8)*, and showed that it could cut DNA at specific positions in the genome that contained the base sequence AAGCTT. Soon, many more restriction enzymes were identified targeting different sequences. Around the same time enzymes performing the opposite reaction, that is the ligation of two DNA molecules, were discovered (9). During this productive period it was also discovered that DNA could be produced from RNA, a process that was named reverse transcription (10, 11), which is a main tool in sequencing of RNA molecules today.

By the mid 1970s it was thus possible to cut and paste DNA, copy it to RNA and visualize it on a gel and identify specific DNA fragments. These methods to manipulate nucleic acids naturally led to experiments where scientists wanted to artificially change DNA in cells, and animals. The first recombinant DNA, that is DNA artificially formed by combining DNA from different species, was created in

1972, by inserting the E. coli galactose operon into SV40 viral DNA (12). This technique was soon applied to create recombinant plasmids that were inserted into bacteria, thereby forming the first recombinant organism (13). The following year, DNA was inserted into a murine blastocyst. The DNA integrated into the embryo genome, and after implantation into a pseudopregnant surrogate mother a healthy recombinant animal was born (14).

Another breakthrough in the 1970s was the first DNA sequencing methods, independently invented by Maxam and Gilbert (15) and Sanger (16, 17). Sanger sequencing is still the golden standard to verify discoveries made by other methods. The method is based on chain-terminating inhibitors. DNA is sequenced by adding an oligonucleotide primer to the template and then allowing the primer to extend in a mixture of normal deoxynucleotides and a much smaller amount of a specific dideoxynucleotide (ddN, N symbolizes any of the bases A, C, G or T) that terminate the elongation when incorporated. If the dideoxynucleotide is ddA, then for each A position in the template some extending strands will terminate their reactions, creating oligonucleotides of different length. The difference in extension length can then be visualized on a polyacrylamide gel. When repeated for all four bases the sequence of the template molecule can be read just by examining the gel.

A great improvement to the molecular toolbox came in 1985 with the development of polymerase chain reaction (PCR) amplification (18). This allowed the amplification of a single DNA molecule to in theory unlimited amounts of copies of the molecule with the help of two oligonucleotide primers and a DNA polymerase. In PCR a double-stranded DNA molecule is denatured, then one primer hybridizes to the (+) strand and the other primer hybridizes to the (-) strand. Both primers are extended leading to two copies of the original double-stranded DNA molecule. This process is repeated until enough copies have been generated for the application of interest.

With Sanger sequencing and PCR amplification at hand leading scientists started to advocate the sequencing of the human genome, a milestone that was completed in 2001 (19, 20). The Human Genome Project (HGP) took 11 years to complete and was achieved with Sanger sequencing. This was the first “Big Science” project in the history of molecular biology and a number of large-scale projects have followed since then.

New exciting tools in molecular biology keep emerging. For example it is now possible to control neuronal activation by light. When deciphering the intricate networks of the brain it is important to be able to turn on or off specific neurons, or types of neurons, with temporal precision (21). This was first tested in 2005 (22) and later termed optogenetics. Boyden et al inserted a protein called

Channelrhodopsin 2 (ChR2) into cultured rat neurons. ChR2 is a transmembrane protein that responds to blue light by a conformational change that opens up a pore in the membrane and lets cations enter the cell. It was shown that pulses of blue light indeed could induce depolarization (neuronal activation). Optogenetics is now a commonly used tool in neurobiology.

One tool that has captured the imagination of scientists, media and laymen alike, is the Clustered Regularly Interspaced Short Palindromic Repeats – CRISPR-associated protein-9 nuclease (CRISPR-CAS9) system for genome engineering. Originally used as a defence system against foreign pathogens in bacteria, the same mechanism is now be used to disrupt or insert genomic content at specific locations in a genome (23). The power of CRISPR-CAS9 stems from its programmability, making it possible to target specific chromosomal locations at will, with nearly single-nucleotide precision.

1.2 NEXT GENERATION SEQUENCING

Upon completion of the HGP the stage was set for Next Generation Sequencing (NGS). Five reasons for the development have been identified. First, Sanger sequencing had been optimized to the point that further optimizations were unlikely to significantly improve cost or throughput. Second, the availability of the human genome made short-read sequencing much more powerful. It was now possible to use the known human genome to map short sequences to their correct places in the genome. Third, many molecular methods had been developed that could benefit from high-throughput sequencing, like RNA-expression and protein-DNA interactions. Fourth, technological development across a number of fields (e.g. microscopy, surface chemistry and polymerase engineering) made alternative strategies for DNA sequencing more feasible (24). Added to this list should be the increase of computing power to handle the massive amount of data generated.

A number of new sequencing methods were developed at this time and have been excellently reviewed elsewhere (24). Here I would like to focus on the methods used in my work: Sequencing By Synthesis (SBS) here demonstrated by the Illumina approach and Single Molecule Real Time (SMRT) sequencing as demonstrated by Pacific Biosciences (PacBio).

In Illumina DNA sequencing, a genome is first fragmented into smaller molecules either by physical shearing or enzymatic cleavage. The ends of the molecules are then repaired enzymatically to form complete double-stranded templates. Adapters are ligated to the templates, and template molecules are subsequently denatured. The single-stranded templates are then hybridized to an array

consisting of short molecules complementary to parts of the adapters. These short molecules function as primers in a PCR reaction and template molecules amplify in a reaction called bridge amplification, forming dense clusters of amplified copies of the template. The generation of clusters is necessary to increase signal to noise ratio when sequencing. Molecules are then made single-stranded again and a sequencing primer is hybridized to the template clusters to a universal region on the adapter flanking the region of interest. Primers are extended one base at a time by nucleotides modified to function as a reversible terminator, i.e. the 3' end is modified not to allow incorporation of additional nucleotides. The modification is reversible and can be removed in a cleavage reaction. Each different type of nucleotide also has a removable fluorophore attached with a specific colour. A sequencing cycle consists of adding all four modified nucleotides, allow base extension by a polymerase, image to capture the fluorescence of the added nucleotide and finally removal of fluorescent dye and the terminator to allow the next cycle of base extension (24, 25). Illumina sequencing is the most successful NGS method so far and has a large part of the sequencing market. However it has limitations in terms of read length where the maximum to date is 700 bp.

The SMRT sequencing method is very different from Illumina sequencing. It uses an aluminium block consisting of many small wells attached to a glass surface. Each well has a polymerase enzyme attached, and single circularized DNA molecules are stochastically added so that each well gets a single molecule based on Poisson statistics. Similar to Illumina sequencing a primer is used to start the sequencing reaction. Nucleotides with a fluorescent molecule attached are added to the wells, and the incorporation of a nucleotide to the extending primer results in cleavage of the fluorophore. Because nucleotide incorporation is slow relative to the diffusion rate, this event can be detected and used for base calling. Thus, the fluorophore from a single incorporated nucleotide, which remains in the well for hundreds of milliseconds, can be distinguished from fluorophore of freely diffusing nucleotides, which shuttle in and out at sub-millisecond timescales. To successfully image a single nucleotide incorporation event from a single polymerase and identify the increase in fluorescence, the reaction volume must be very small. This is achieved by making the wells in the aluminium plate so small that the reaction volume is measured in zeptoliters (10^{-21} liter) (26). The main benefit of PacBio sequencing is the read length that can cover several kb, however the throughput is limited compared to Illumina sequencing.

Next generation sequencing has made possible a number of large-scale projects. A common aim of these projects is to systematically and in depth examine questions of particular importance in molecular biology, typically generating a massive amount of data to allow computational and mathematical modelling to provide answers to the question. This is sometimes referred to as systems

biology. A typical example is The Cancer Genome Atlas that catalogues genetic mutations responsible for cancer. Other examples include the Encyclopedia of DNA Elements (ENCODE), which aims to build a list of functional elements in the human genome. One controversial finding of ENCODE is that most of the human genome, around 80%, are in some way active in at least one cell type (27), and it's implicated that a large part of that is also functional. A serious critique of this finding is that only around 10% of the genome is evolutionarily conserved (28), implying that the majority of the genome is expendable and probably not functional. Another large-scale project is the Functional annotation of the mammalian genome (FANTOM) project, which aims to assign functional annotation to full-length cDNA. Recently the FANTOM project mapped transcription start sites (TSS) by their Cap Analysis of Gene Expression (CAGE) method, and data from this effort has been used in paper III.

1.3 THE NEED FOR ENHANCED PRECISION IN SEQUENCING

The ideal sequencing method should be able to take a tube filled with nucleic acids and translate all molecules in the tube into correct sequences in the computer. This is not possible today because all sequencing methods introduce bias and errors when converting the chemical molecules into digital information. Also sequencing instruments require that input molecules comes in a certain form and quantity. This transformation of genomic DNA or RNA into molecules suitable for a certain instrument is called library preparation, and this processing also introduces bias and errors. An especially biased and error prone step during library preparation is the amplification step, which is necessary in many protocols.

For many applications it is important to distinguish artefactual mutations from the library preparation and sequencing processes, from real mutations in the genome. Consider the case of monitoring residual disease in leukaemia. The ability to identify early on the type of mutation that renders a cancer clone resistant to therapy can be the difference between life and death. However if library preparation and sequencing introduce errors of a magnitude of 1 error per 100 bases sequenced, mutations with a frequency of less than 1% will not be accurately called. This means that a relapsing clone already has a substantial amount of copies before it can be identified and addressed.

Tools have been developed to greatly improve quantification and decrease errors in sequencing data. Part 2 of this thesis discuss why these tools have been developed, how they work and in what areas of molecular biology they have been applied.

2 COUNTING AND ACCURATELY SEQUENCING NUCLEIC ACIDS

2.1 INTRODUCTION

There is a need for more accurate and quantitative measurements of nucleic acids than what is possible with the sequencing technologies available today, especially in the biomedical areas. This need is accentuated in cases where extensive amplification is needed, such as single cell sequencing, targeted sequencing of rare mutations or sequencing of clinical material where only very small amounts are available.

The methods developed to increase accuracy of sequencing data involve marking individual molecules with a molecular barcode and thereby making them unique. This allows both for accurate quantification and error correction. Since previously identical molecules are now unique, it is possible to count the molecules even after amplification, and by comparing multiple copies belonging to a single molecule it is possible to identify artefactual mutations coming both from clonal amplification errors and random errors produced by sequencing.

Part 2 of this thesis describes why it is important use molecular barcoding of individual molecules, how the method was developed, and the different strategies used to improve quantification and error correction of NGS data. It will also show a number of examples where the barcoding strategy has been used and compare the different barcoding strategies. Finally it summarizes the developments made to the barcoding strategy and discusses pitfalls that should be avoided.

2.1.1 Amplification-induced bias

Illumina sequencing is the most widely used sequencing technology today, mainly due to its high throughput and accuracy. Illumina sequencing usually requires amplification during library preparation and then another round of solid-phase amplification during the sequencing reactions. Amplification is always uneven and introduces errors, although substantial improvements have been achieved in polymerase efficiency and fidelity. For example AccuPrime Pfx has been shown to have an error rate of 2.9×10^{-6} (29) and KAPA HiFi claims an error rate of 2.8×10^{-7} (30). Some molecules, usually short molecules with a normal GC (Guanine - Cytosine) content and no secondary structures, amplify more efficiently than others. The extent of this bias varies substantially with the

polymerase used (31). Interestingly a recent study using AccuPrime Pfx showed with both theoretical modeling and experimental evidence that the main source of PCR error is stochasticity of amplification in the first rounds of PCR amplification, followed by polymerase errors. GC content and template switching had only minor influence (32).

Amplification creates four problems as illustrated in figure 1. The first problem regards quantification. If X DNA molecules with the exact same base pair sequence are amplified, and the amplification efficiency is unknown, then information about *X* is lost. The second problem has to do with introduction of errors. Amplification introduces errors depending on factors such as the number of cycles of amplification, polymerase fidelity and read length. Errors consist of base pair substitutions, deletions or insertions, or in the form of template switching where the final PCR product consists of a combination of two or more molecules. The third problem has to do with redundant information. Sequencing data from a highly amplified library will contain many copies of transcripts that don't provide additional information. In other words the yield, defined as the number of informative bases divided by the total number of bases, is low. The fourth problem has to do with loss of information. Poorly amplified molecules will be sequenced to a much lesser extent, e.g. leaving large areas with aberrant GC content unsequenced in genomic sequencing. In conclusion, amplification doesn't only confuse the analysis by introducing errors and bias; it also limits sequencing capabilities by repeatedly sequencing the same molecules while leaving out regions from being analysed.

A couple of methods have been developed to deal with the quantification problem. For example, a patient sample and a healthy control can be analysed simultaneously and since the amplification efficiency, which depends on GC content, length and secondary structures of the DNA, should be similar in the sample and control, it is possible to get an estimation of the relative abundance of gene transcripts in the sample. It is not possible to determine the number of unique molecules, but it is possible to determine if there is an overall increase or decrease in gene expression in the sample compared to the control.

It can be very important to get an exact quantification of the number of identical molecules in a sample, e.g. to determine the number of cells with a mutant allele in cancer diagnostics or in monitoring of residual disease. Digital PCR was developed to this end (33). In digital PCR DNA is diluted to a low concentration and divided into wells in a plate so that each well contains only one copy of the template molecule. Each molecule is then amplified and detected. This allows an exact quantification of the number of participating molecules, however the throughput is low, allowing only for querying a limited number of targets at a time (34).

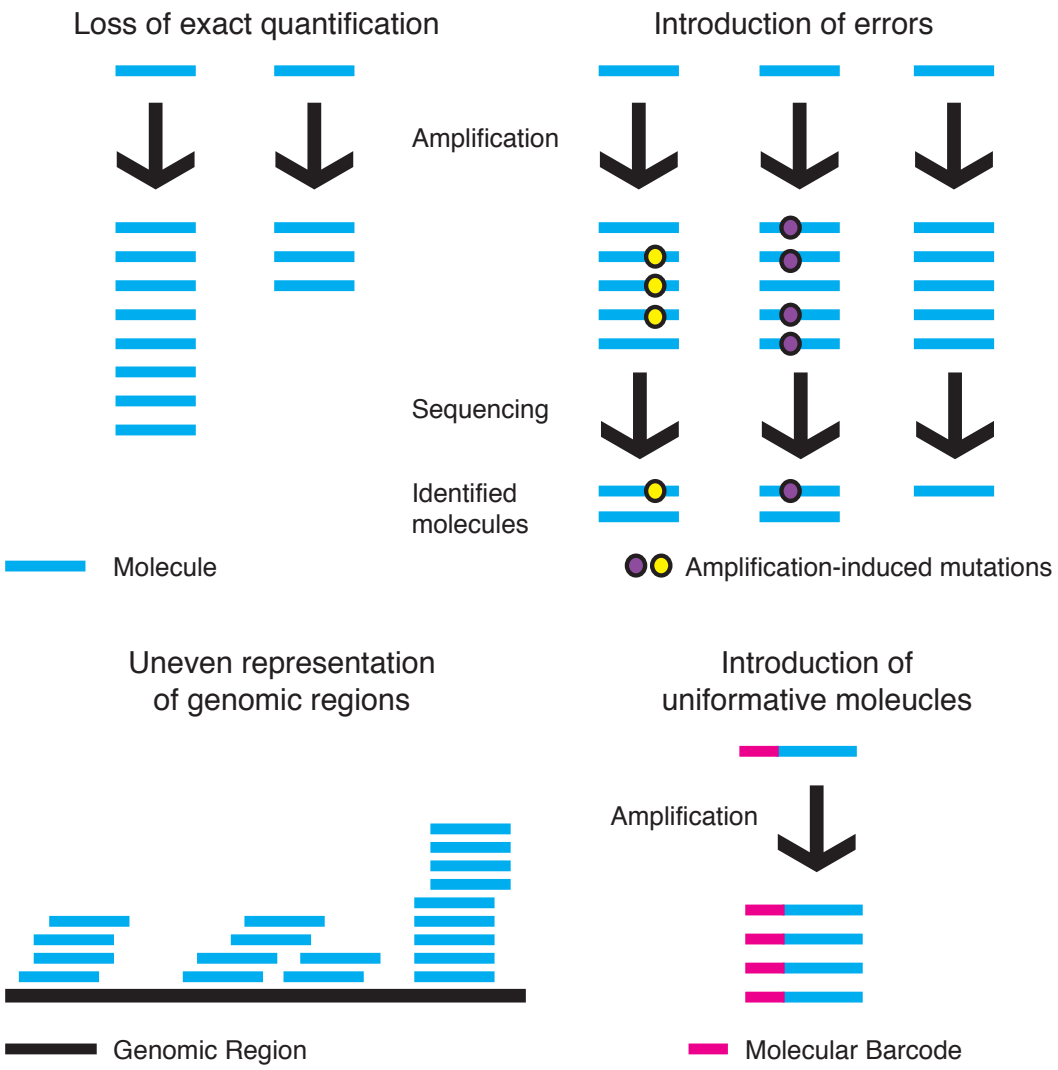

Figure 1: Problems related to amplification of molecules before sequencing.
Four problems are created by the amplification process: loss of quantification, introduction of errors, uneven representation of genomic regions and introduction of uninformative molecules.

2.1.2 Unique Molecular Identifiers

Unique Molecular Identifiers (UMI) was introduced as a measure to solve the quantification and error-introduction problems of amplification. A UMI is a barcode that is attached to the template molecule in an early step of the library preparation. The barcodes make each molecule unique before amplification. What does this mean? A simple analogy would be the problem of retrieving your traveling bag after a flight. Often people waiting at the baggage carousel have the exact same brand of traveling bag and it is impossible to tell them apart. Your simple black bag is not possible to tell apart from the other black bags going around on the carousel. To solve this problem savvy travellers have begun to

mark their bags with colourful belts or silly stickers, and that is exactly what the UMI does. The classical approach is to ligate a barcode consisting of a number of random nucleotides. If the length of the barcode is sufficient in relation to the number of identical molecules in the sample then the chance is low that two identical barcodes ligate to two identical molecules and each molecule therefore becomes unique. When each molecule is unique, quantitative information will not be lost by amplification.

Molecular barcodes can solve the quantification problem and correct for errors created by amplification, however they cannot identify transcripts that amplify poorly and each barcode is often sequenced multiple times resulting in many non-informative sequence reads. To sequence genomic regions that amplify poorly or to sequence only new molecules an amplification-free method should be used (35).

Molecular barcoding of individual DNA molecules to improve quantification was first proposed in 2003 (36). Already in 2004 the theory was put into practice for the first time (37), and this was followed up and improved by the same group in 2007 (38). In 2010 Konig et al also used the molecular barcoding strategy to discriminate between unique DNA sequences and PCR duplicates (39). Using molecular barcodes to create a consensus sequence for error correction was introduced soon thereafter (34, 40-43). By 2012 the concept was well established and had been used in a number of different settings, among these amplicon sequencing of a viral RNA, identification of mutation frequencies in a small genomic region and karyotyping of cell-free DNA. The coming years saw a number of tweaks and improvements to the method and by now it has been referred to in more than 100 scientific papers.

2.1.3 Turning a quantitative problem into a qualitative problem

One of the first uses of the concept of molecular barcoding were in signature-tagged mutagenesis (44, 45), to track the origins of expressed sequence tags (what is now be called multiplexing) (46) and in labelling objects with DNA for identification (47, 48).

The first paper to mention molecular barcoding of individual molecules to reduce amplification-induced bias was a theoretical paper by Hug and Schuler in 2003 (36). They proposed turning the quantitative problem of counting exact number of molecules of a single mRNA species in complex soup of mRNA molecules, into a qualitative problem by making each individual molecule unique. Another way to describe the solution is to contrast it to digital PCR: In

digital PCR molecules are separated in the physical space, in molecular barcoding molecules are separated in the chemical space (49).

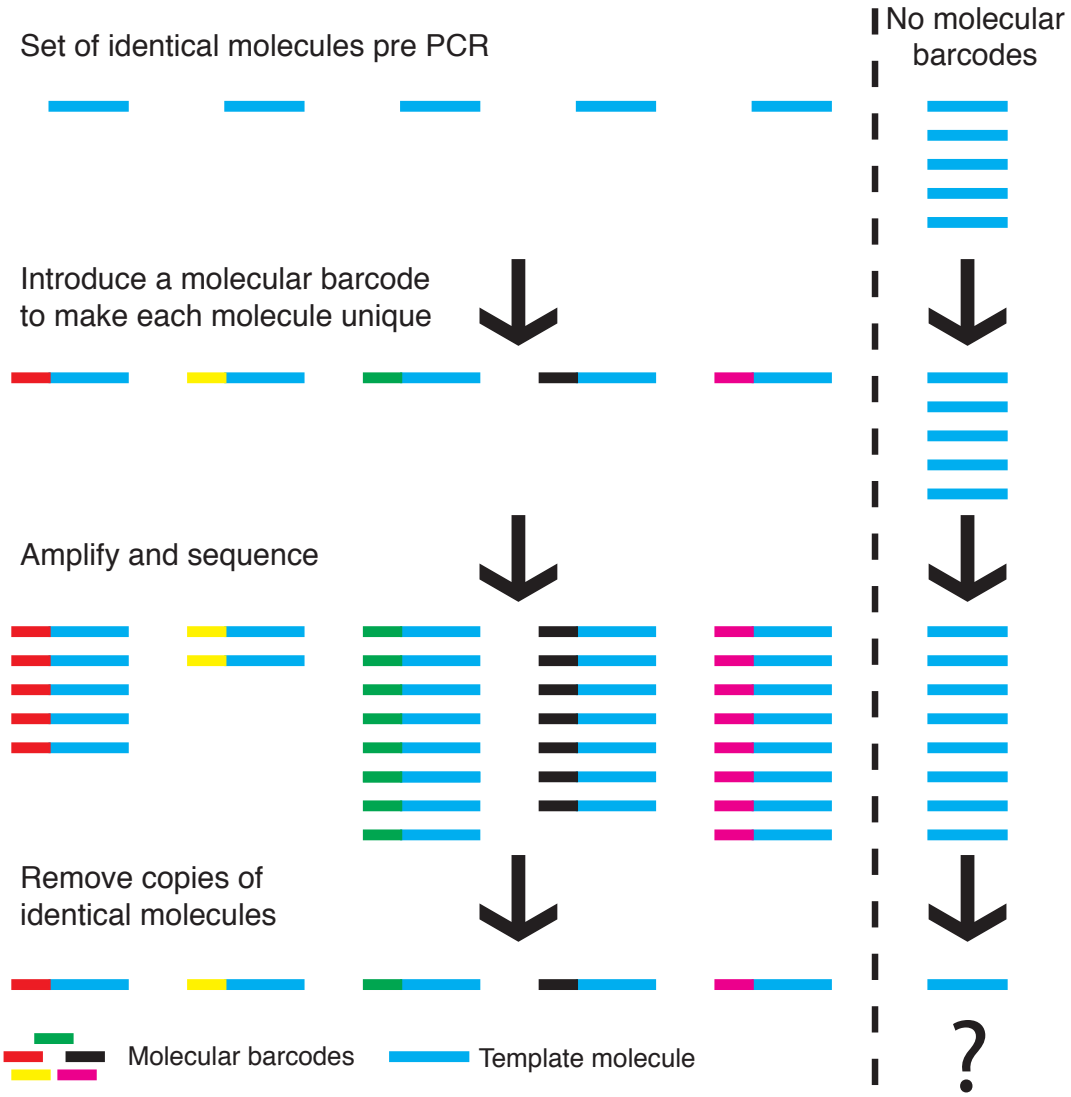

Figure 2: Molecular barcoding of individual molecules to correct for quantification bias. Comparison between amplification of molecules with and without a molecular barcode. Here five identical molecules are amplified. To the left, amplification with a barcode. To the right, amplification without a barcode. When a library with barcodes is amplified it is easy to keep track of the number of original molecules, provided that each molecule is sequenced at least once. When amplification is done without barcodes information about the number of original pre-amplified molecules is lost.

The method was suggested to comprise of four steps: First to isolate the mRNA species of interest, second to make each mRNA molecule unique, into something they called Different Molecular Species (DMS), third to amplify the DMS and fourth to detect and count the DMS (outlined in figure 2). Thereby the quantitative problem of counting mRNA molecules of a certain species was transformed into the qualitative problem of detecting DMS. The qualitative

problem is solvable just by sequencing deeper, that is the sensitivity for picking up new DMS increases with sequencing depth, and thereby the resolution in counting molecules improves. This is in contrast to the quantitative problem where increased sequencing depth doesn't improve counting resolution, since additional copies of the same molecule don't add more information.

2.1.4 Putting theory into practice

In 2004 the method was invented again and this time it was applied to bisulphite sequencing of the *FMR1* promoter region in the DNA of males with fragile X syndrome (37). The molecular barcode consisted of a degenerate sequence absent of cytosine, since the identity would be ambiguous after bisulphite conversion. From the 8 samples in their study they identified redundant sequences ranging from 7-51% and that number appeared to be influenced by the amount of input DNA.

2.2 DEVELOPMENT OF MOLECULAR BARCODING STRATEGIES

Already in 2004 molecular barcoding of individual molecules had been theoretically explained and practically tested. Still the method didn't get appreciated by a broader audience until 2011. By then NGS was well established and this had accentuated the need for accurate quantification and error correction.

2.2.1 Molecular barcoding for error correction

Casbon et al examined the performance of the molecular barcode (here called molecular counter) in amplicon sequencing (40) and showed that the molecular counter substantially lowered allelic bias, and could be used for error correction. They digested human genomic DNA with a restriction enzyme and ligated adapters containing random barcodes, here called degenerate base regions (DBR). Molecules were circularized, PCR amplified by inverse PCR and sequenced on the Roche 454 platform.

To test the molecular counter, different amounts of input DNA were used in the inverse PCR reaction (50, 100 and 250 ng). Unsurprisingly there was no correlation between the number of reads and the input mass (R^2 = -0.20), yet there was a correlation between input mass and the number of molecular counters (R^2 = 0.78), showing that the counter was sensitive to an increased amount of input DNA. Most DBRs identified were singletons, which is when a molecular barcode is only supported by a single read, but some DBR had high

read numbers, up to 455 reads for the 50 ng input sample, an indication of imbalanced amplification.

Molecular barcoding is also useful for error correction. Most amplification-induced errors occur in the later stages of amplification due to the increased number of DNA molecules present, allowing more opportunities for polymerase errors. Provided that a template molecule is sequenced to sufficient depth, these late-occurring errors will constitute a minor part of all sequences with the same molecular barcode. By using a consensus nucleotide for each position in the template molecule such errors can be removed (shown in figure 3). Casbon et al were able to show that by using a consensus read from molecular counters the error rate, in terms of artefactual mutations, insertions and deletions, was substantially lowered compared to using reads.

Added to the study was a theoretical discussion about the likelihood of collision events. A collision occurs when two different identical molecules receives an identical barcode by chance. This is more likely to happen if the input DNA is limited and the number of bases in the degenerate sequence are few. Through the use of Bayes theorem the authors conclude that the number of collisions will be few when the number of template molecules are fewer than or equal to the square root of the number of possible combinations of the degenerate barcode. So if you expect to interrogate 100 identical molecules a molecular barcode of complexity 10'000 should be used, e.g. a degenerate barcode with the size of at least seven (4^7 = 16384). The authors also note that it is more important to get an exact quantification when the template molecules are few, i.e. it is more important to be able to quantify the difference between 1 and 5 molecules than between 51 and 55.

2.2.2 Redundant sequencing for improved error correction

Redundant sequencing is important both to ensure that most or all molecules are sequenced and for allowing the creation of a consensus sequences. Kinde et al developed a molecular barcoding strategy where this was an important ingredient (41). Their method, called the Safe-Sequencing System (Safe-SeqS), comprised three steps: (i) assignment of a unique identifier (UID), (ii) amplification of each uniquely tagged template to create UID families, (iii) redundant sequencing of the amplification products. Safe-SeqS had a read depth averaging 15-107 reads per molecule and used a minimum 2 reads to construct a consensus sequence. Examined mutations on PCR fragments with the same UID were considered a true mutation only if more than 95% of them contained the identical mutation, here called a supermutant.

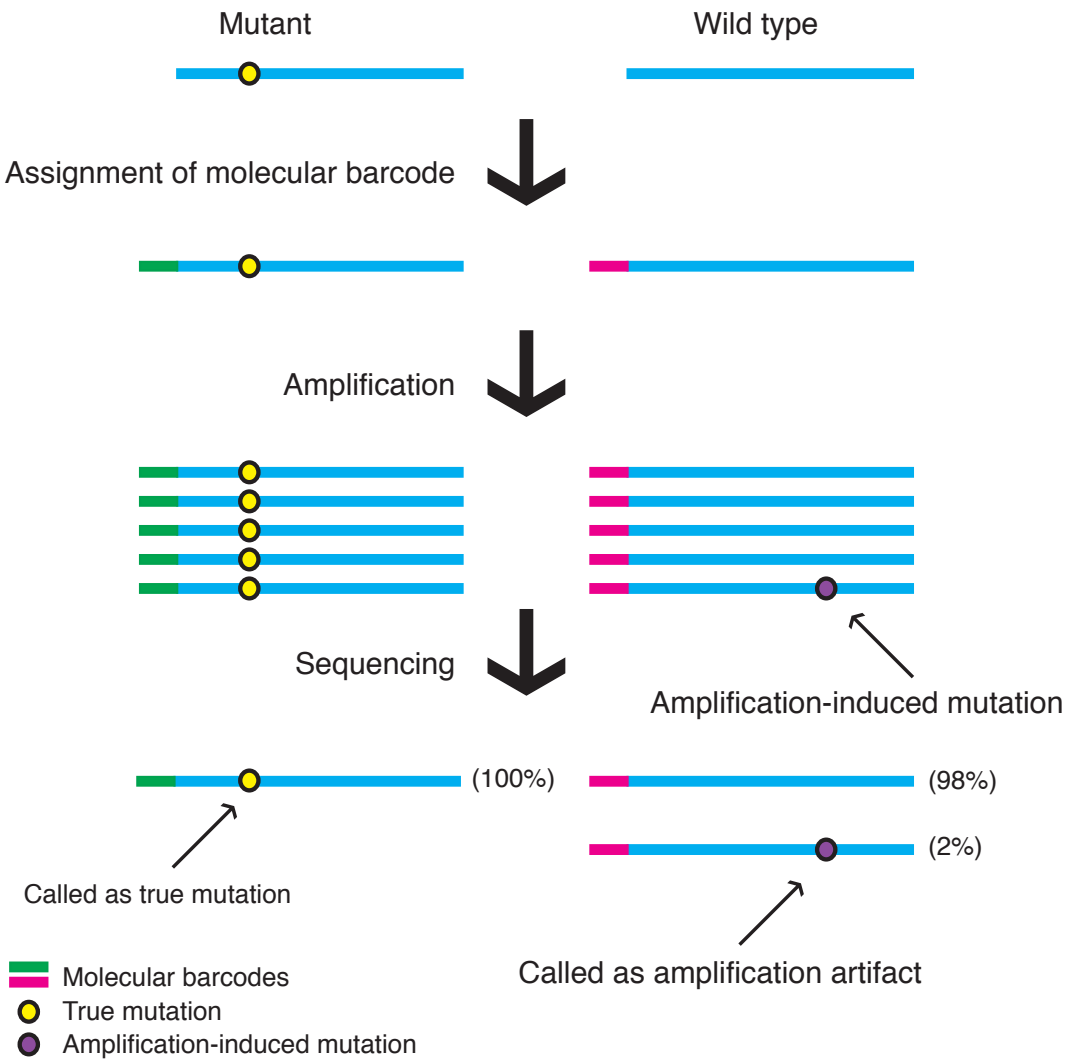

Figure 3: Molecular barcoding of individual molecules for error correction.
Molecular barcodes can be used to correct for amplification or sequencing induced errors. A true mutation should be present in all or most copies in a group of molecules with the same barcode. Amplification-induced errors are more common in the later stages of PCR and each such error normally constitutes a minority of all molecules with identical barcodes.

In Safe-Seqs genomic DNA amplicon targets were amplified two cycles with amplicon-specific primers where both primers contained a universal tag sequence and the forward primer contained a random barcode of 12-14 nucleotides. The tagged templates were then amplified an additional 25 round with primers hybridizing to the universal sequences and then sequenced on an Illumina GA IIx instrument.

Safe-SeqS was applied to study the prevalence of rare mutations in a small region of the CTNNB1 gene isolated from ~100'000 humans cells from three unrelated individuals. Conventional analysis of Illumina sequencing resulted in a mutation rate of around 2×10^{-4} mutations per bp. Safe-Seqs however had a mutation rate of 9×10^{-6} mutations per bp, reducing the frequency of mutations in genomic DNA

24-fold, confirming that most mutations identified without the molecular barcode were artefactual.

Kinde et al made a contribution to the use of molecular barcodes by introducing the concept of endogenous versus exogenous barcodes. An exogenous barcode is a barcode that is added to the molecule of interest e.g. through adaptor ligation or limited cycles of PCR during library preparation. The exogenous barcode can in principle be made to contain an unlimited number of different barcodes. If the barcode contains any of the A, C, G or T bases then the limit is in how long the sequence is and the number of possible barcodes will be 4^N. Kinde et al realised that it was possible to use the ends of molecules that had been randomly sheared as the UID (illustrated in figure 4). If shearing occurs randomly it is unlikely that a fragment will have the exact 5' and 3' end as another fragment. This endogenous UID doesn't require an externally added barcode, but has a limited number of unique ID's, and is suitable when the number of targets is limited. Here Kinde et al showed that it was possible to use the endogenous UID when studying DNA mutations from around 15'000 cells.

2.2.3 Mutational hotspots to further reduce errors

Building on to the approach taken by Kinde et al, Shugay et al developed a method to further reduce artefactual mutations (50), named Molecular Identifier Groups-based Error Correction (MIGEC).

MIGEC exploits the phenomenon that most artefactual mutations are produced in the later stages of PCR amplification and are therefore filtered out when the consensus sequence is constructed. The artefactual mutations are often reproducible and Shugay et al showed that for a specific sequence >90% of all errors were reproducible and therefore mutational "hotspots" could be identified, as shown in figure 5. On top of error correction based on the consensus sequence, in MIGEC these mutational hotspots are also filtered out.

It was also observed that due to amplification-induced errors within the UMI sequence itself, the amount of UMIs was inflated (51). A 12 bp long UMI amplified and sequenced 10^4 times often resulted in 10-20 erroneous UMI subvariants due to mutations in the UMI sequence. If no correction was made this would result in an estimation of 11-21 template molecules when the true number of molecules was 1. Since most UMI errors are created by a single mismatch from the true UMI and have a low read coverage, a two stage UMI filtering approach was applied in MIGEC where minor UMI subvariants that differed by a single nucleotide were removed and then a threshold was set of at

least several sequencing reads per UMI. The optimal threshold varied depending on the size of the starting library and the level of over-sequencing achieved.

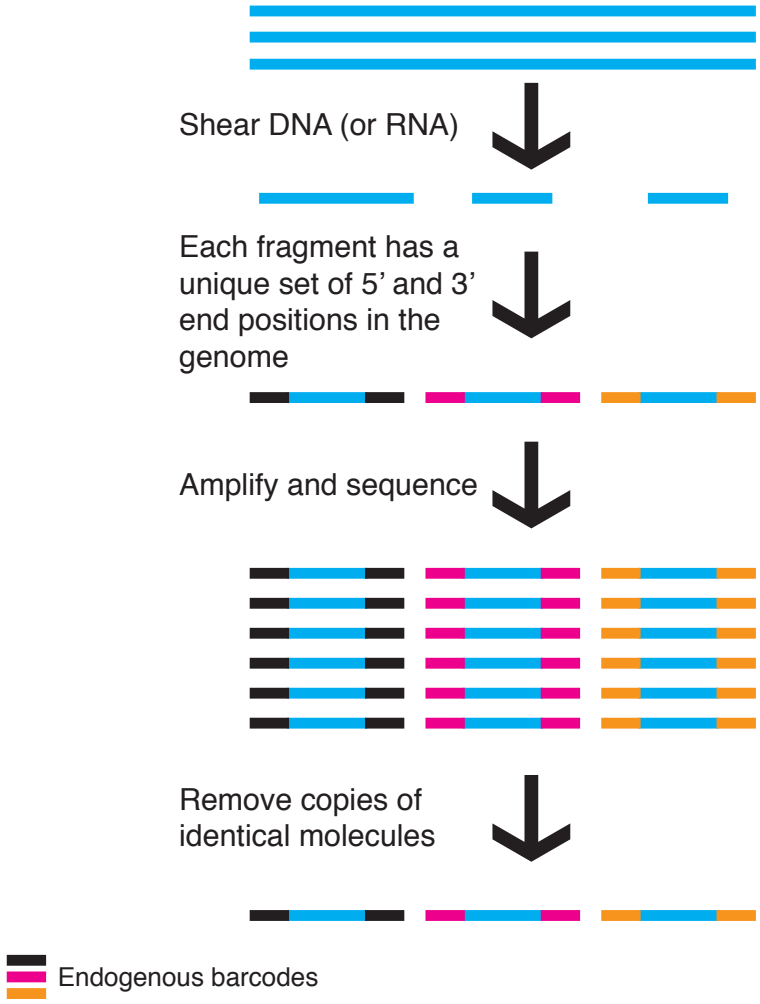

Figure 4: Endogenous barcodes
Barcodes can be prepared by random shearing of DNA or RNA. If molecules are sufficiently diluted and randomly sheared each molecule will have a unique start and/or end position in the genome, and is therefore unique with a high probability in a similar manner as the exogenous barcode.

2.2.4 Error correction by the complementary strand to remove artefacts from the first round of amplification

The methods described above all show remarkable abilities to reduce amplification and sequencing induced errors. But none of them were able to handle errors introduced in the first round of amplification. MIGEC can correct for hotspot mutational errors in the first round of amplification, but not errors coming from some other sources. Therefore a certain uncertainty existed about how prevalent these errors were.

To correct such errors a new method was developed called Duplex Sequencing. It reduces artefactual amplification-induced mutations by using the double-stranded nature of DNA for error correction (43). If a mutation arises it should arise in both strands, and if both strands are sequenced then the mutation should exist in both strands in the sequencing data, otherwise the mutation is an artefact. A benefit of this method is that it captures errors also in the first round of amplification, which frequently happens when the DNA is damaged or degraded. Schmitt et al argues that since the mutation rate during cell division is estimated to range from 10^{-8} to 10^{-11} per nucleotide, the majority of the mutations called with the previously mentioned error reduction methods still potentially represent technical artefacts. Also, commonly used DNA polymerases for library construction have a misinsertion frequency rate between 10^{-4} and 10^{-6} (52), which leads to many false positives even in the first round of PCR amplification. In order for Duplex Sequencing to interpret an artefactual mutation as a "true" mutation, the complementary artefactual mutation must occur on the complementary strand as well, which is a very unlikely event.

In Duplex Sequencing, adapters are created by two partially complementary oligonucleotides, one of which has a 12 bp degenerate overhang. The overhang was rendered double stranded by a DNA polymerase and then adapter A-tailing was performed by incubation of the adaptors with DNA polymerase and dATP (Deoxyadenosine triphosphate). Genomic DNA is prepared with shearing and end-repair by standard protocols. A T-overhang is created by polymerase elongation of template DNA with dTTP (Deoxythymidine triphosphate). Adapters are ligated to the template and the product is PCR amplified and sequenced with paired-end sequencing. Reads are grouped based on the UMI and consensus sequences are created. Thereafter complementary sequences are identified based on the UMI. Since both ends of the template molecule have a UMI, both ends must match to call a complementary pair. After successful pairing the template sequences are compared and only bases matching perfectly are kept and called duplex consensus sequences (DCSs) as illustrated in figure 6.

To test Duplex Sequencing M13mp2 DNA, a substrate extensively studied and that has a well-established base substitution frequency of $3*10^{-6}$, was sequenced. Standard analysis methods including quality filtering for a Phred score higher than 30 resulted in an error frequency of $3.8*10^{-3}$, more than 1000-fold higher than the true mutation frequency. Consensus filtering based on the UMI resulted in an error frequency of $3.4*10^{-5}$, suggesting it corrected for 99% of all sequencing errors. Still this number was > 10-fold higher than the reference value. When Duplex Sequencing analysis was applied, the mutation frequency went down to $2.5*10^{-6}$, almost identical to the reference value of $3*10^{-6}$.

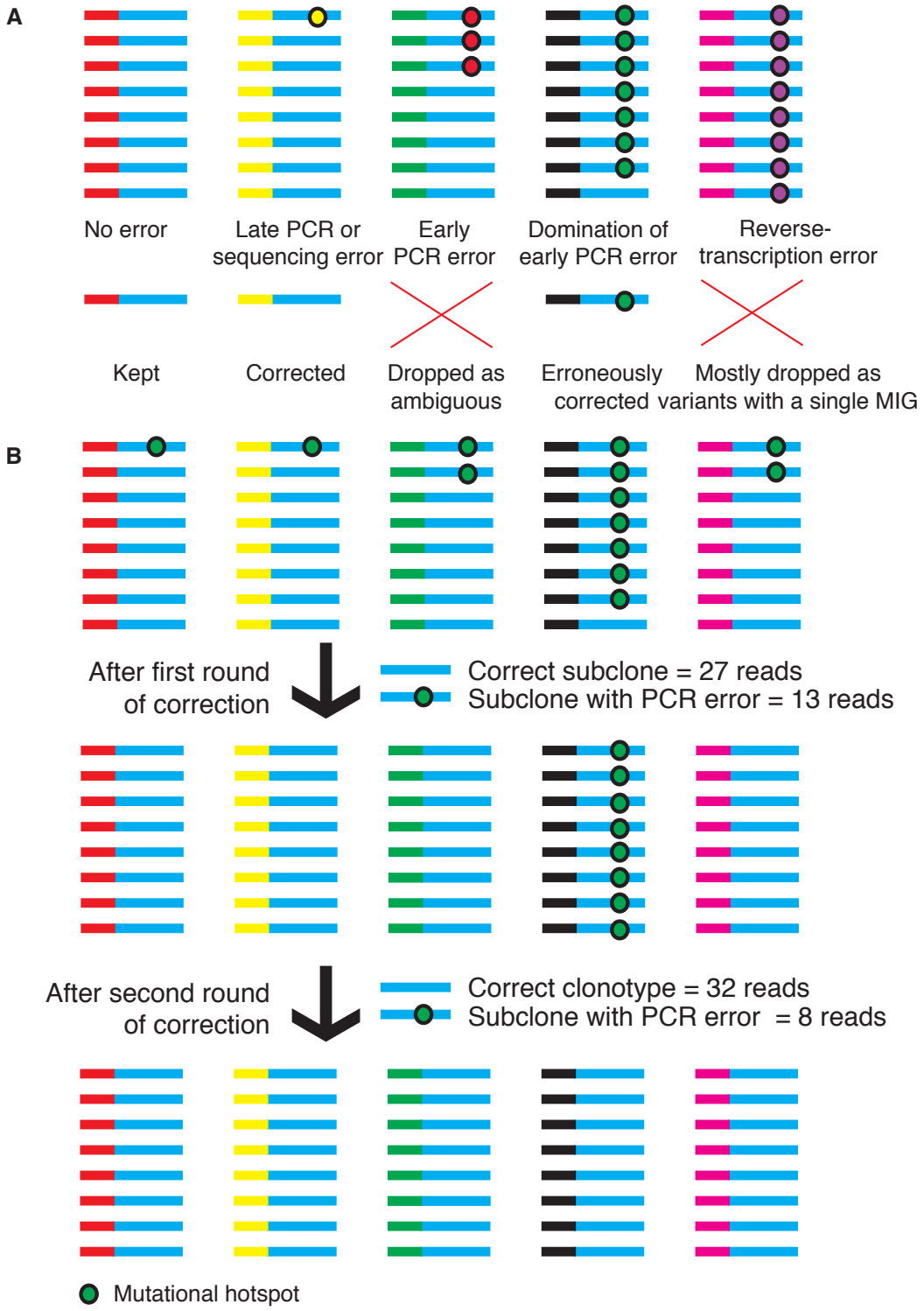

Figure 5: Mutational "hotspot" correction of errors

A) Possible scenarios of error distribution among reads with the same barcode. B) Early amplification-induced error rarely dominates the read family. Late amplification-induced errors are comparatively common, and often occurs at the same position between different read families of the identical molecules. This can be used to identify and correct early polymerase errors. Correct subclones always gain reads after the first correction, while subclones with a repetitive PCR error lose reads. In this example the correct subclone gained 5 reads (32-27) and the erroneous one lost 5 reads (13-8). The position was therefore identified as a mutational hotspot and was corrected in the second round of correction. MIG – Molecular Identifier Group. Adapted from Shugay et al (2014).

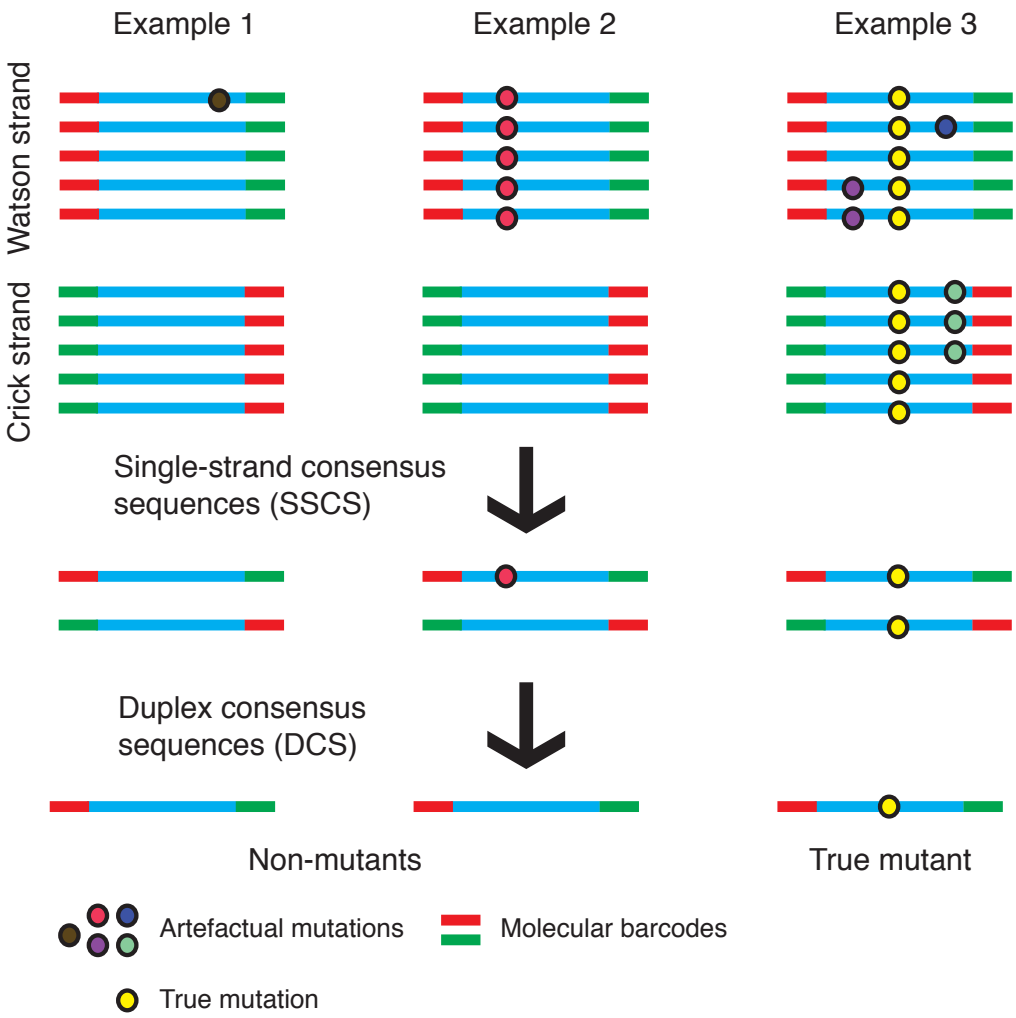

Figure 6: Using the complementary strand to correct for errors – Duplex Sequencing
Three examples of read groups with the same molecular barcode are shown. The first example shows a read group with a random mutation in one strand, which is interpreted as a false mutation. The second example shows a mutation that is consistent on one strand, but is absent on the other; this in interpreted as an artefactual copy of the nucleotide, probably due to DNA damage. The third example shows that a true mutation must exist in both strands. Adapted from Schmitt et al (2012).

Interestingly, errors not corrected for by duplex consensus building had a specific mutational pattern with an excess of C→T and G→T relative to the reference. G→T mutations can be explained by the fact that an A is commonly inserted opposite to 8-oxo-G leading to misinterpretation of G as a T. The G→T was also in great excess of the C→A mutation, which if all mutations had been correct, would have had the same frequency. The C→T mutation is consistent with a spontaneous deamination of cytosine to uracil, also leading to insertion of adenine opposite to the uracil and a misinterpretation of C as a T. These artefactual mutations were not seen in the Duplex Sequencing corrected data.

Schmitt et al suggested that the difference in mutation frequencies between the UMI method and the Duplex Sequencing method could be exploited to study the extent of oxidative DNA damage in a sample. Duplex Sequencing has also been used to estimate the mutational frequency of human brain mitochondrial DNA (53), and it was used to detect rare mutations in the ABL1 gene that confer resistance to imatinib when treating chronic myeloid leukaemia (54)

2.2.5 Error tolerant barcode set to account for UMI mutations

Making the molecular barcode degenerate is a convenient way of preparing and handling the adaptor or primer – you only need one barcode. However library preparation easily introduces errors to the barcode leading to the introduction of artefactual unique molecules, and at least some, if not all, of the singleton UMIs can be attributed to this. To overcome this problem a method was invented that use a pre-determined set of barcodes to replace the degenerate barcode. The set consisted of 145 barcode sequences 20 bp long, designed in such a way that each barcode had a Hamming distance of 4, that is they could sustain up to four substitution errors and still be unambiguously identifiable (55). The benefit of this is illustrated in figure 7. To increase the number of labels the barcode was attached to each end of the template molecule and then paired-end sequencing was used to identify both molecular barcodes. In this way 145*145 = 21,025 unique labels was be created. This also has the added benefit that if one end had a strong bias towards a certain barcode, the other end is unlikely to be equally biased.

Shiroguchi et al tested the improved barcoding strategy on a set of spike-in molecules where one sample was labelled with a degenerate barcode and one sample was labelled with the pre-determined barcodes. The comparison between the two barcoding strategies, with a degenerate barcode 16 bp long, showed that the degenerate barcode produced an estimate of pre-PCR molecules 15.5-fold higher than the estimated input of 10’000 molecules. A large part of the degenerate barcode estimation came from barcodes with a low read-count. The over-estimation using degenerate barcodes was lowered to 2.4 fold under the assumption that two molecules with up to two mismatches in the barcode region were originally from the same DNA molecule.

Interestingly the predetermined set of barcodes had very few barcodes with a low number of reads indicating that a large part of degenerate barcodes with few reads are artefacts created during amplification and sequencing. The predetermined set of barcodes underestimated the number of molecules by 0.5 fold.

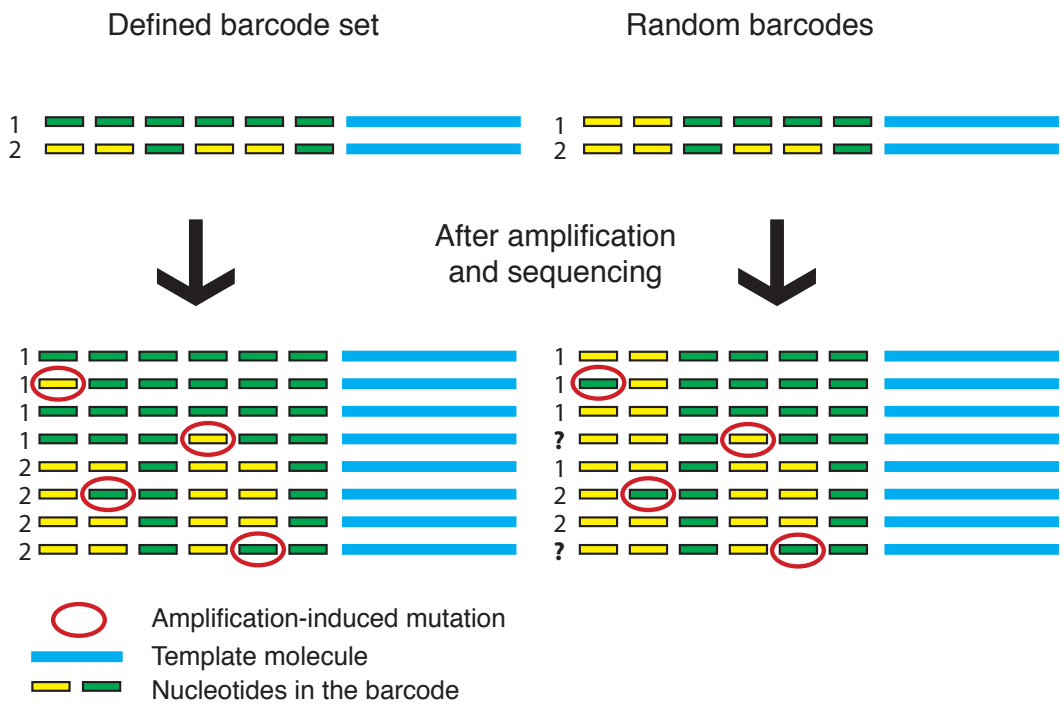

Figure 7: Error tolerant barcode set to identify and correct for UMI mutations
A defined barcode set is tolerant to a certain number of substitution mutations. Here a simplified illustration is shown with only two types of nucleotides instead of four. To the left, two barcodes with a Hamming distance of 4. When a mutation occurs in such a barcode it is still possible to tell which barcode the mutated barcode belonged to, since it will be more similar to that barcode than to any other barcode. To the right, a random barcodes where two molecular barcodes by chance got a Hamming distance of 2. In this case a single mutation can make it impossible to infer which original barcode the mutated barcode belonged to. In reality only the information after amplification and sequencing is known and in this case the barcode sequences on the right could be interpreted as representing 1-6 template molecules depending on how stringent the barcode consensus criteria is set.

2.2.6 Creating limiting dilution using primers instead of template

Individual UMIs should preferentially be sequenced multiple times both in order to facilitate an accurate error correction, and to be confident that all individual molecules have been detected. In many applications this requires that the DNA from each sample is diluted before applying adaptors or primers. It is both cumbersome to measure the DNA concentration of each sample, and it can be difficult to make each sample equally diluted, creating a bias between samples. Faith et al suggests a method, Low-Error Amplicon Sequencing (LEA-Seq) (56), to circumvent this problem: Instead of creating a limited dilution of the samples, the PCR primers in a first round of PCR are diluted, as illustrated in figure 8. This set of diluted PCR primers can then be used on all samples.

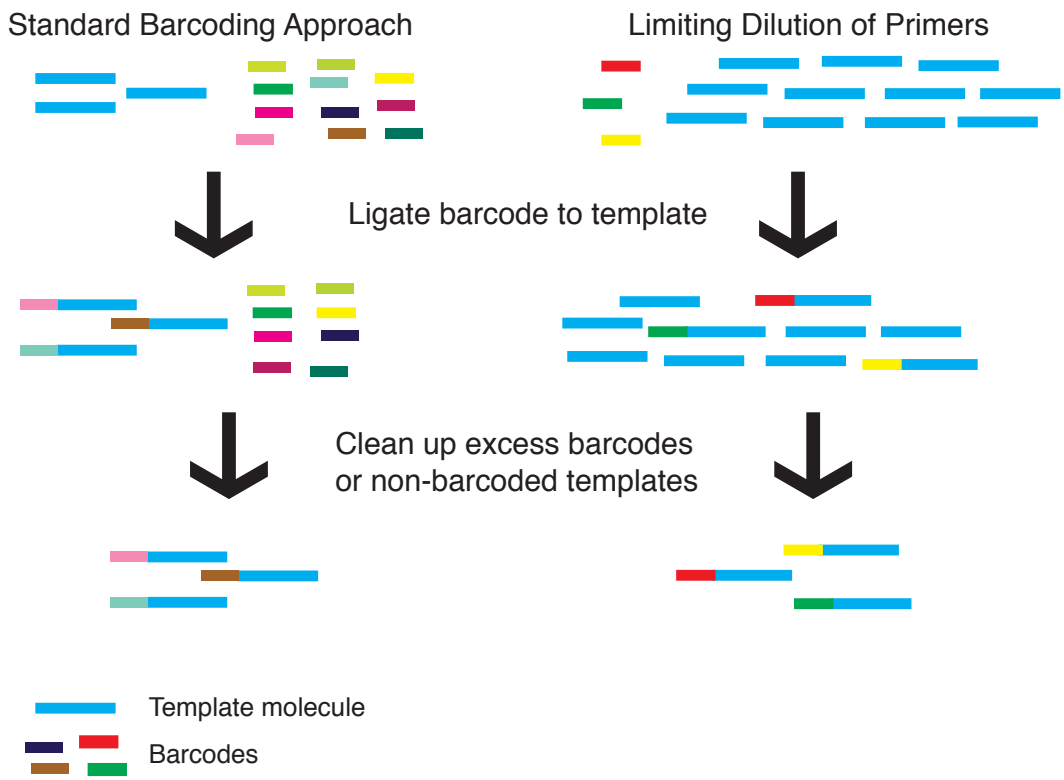

Figure 8: Using primers as limiting dilution.
Comparison between a standard barcoding approach to the left where template molecules are diluted before barcoding and an excess of barcodes are used, and the limiting dilution of primers approach to the right, where the number of primers are diluted before the barcoding reaction and the number of template molecules are left intact.

2.2.7 Singleton UMI molecules

A general problem with molecular barcoding is that often a surprisingly large part of all molecules are singletons. This is due to mutations in the UMI sequences or that the library is not sequenced deep enough in relation to its complexity. The latter can be resolved by sequencing deeper, while the former is harder to work around. If a predefined set of barcodes is used then mutations in the UMI is a rather straightforward problem to solve, but in many applications a degenerate sequence is used.

At least two strategies have been proposed to handle singletons: either they are removed, which can lead to substantial loss of data, or molecules with very similar UMI sequences are grouped together (51), which potentially leads to an underestimation of the true number of UMI. Lundberg et al proposed a third method to deal with singletons (57). They noted that in template-overloaded samples, that is samples with few reads per barcode in average, more singletons will be correct compared to libraries with a high ratio of reads per barcode, as shown in figure 9. In their study sequencing reads were either analysed as

untreated (non-consensus), as consensus sequences, or as Consensus sequences plus Adjusted Singletons (CASs), where singletons were downsampled in proportion to reads per molecular barcode.

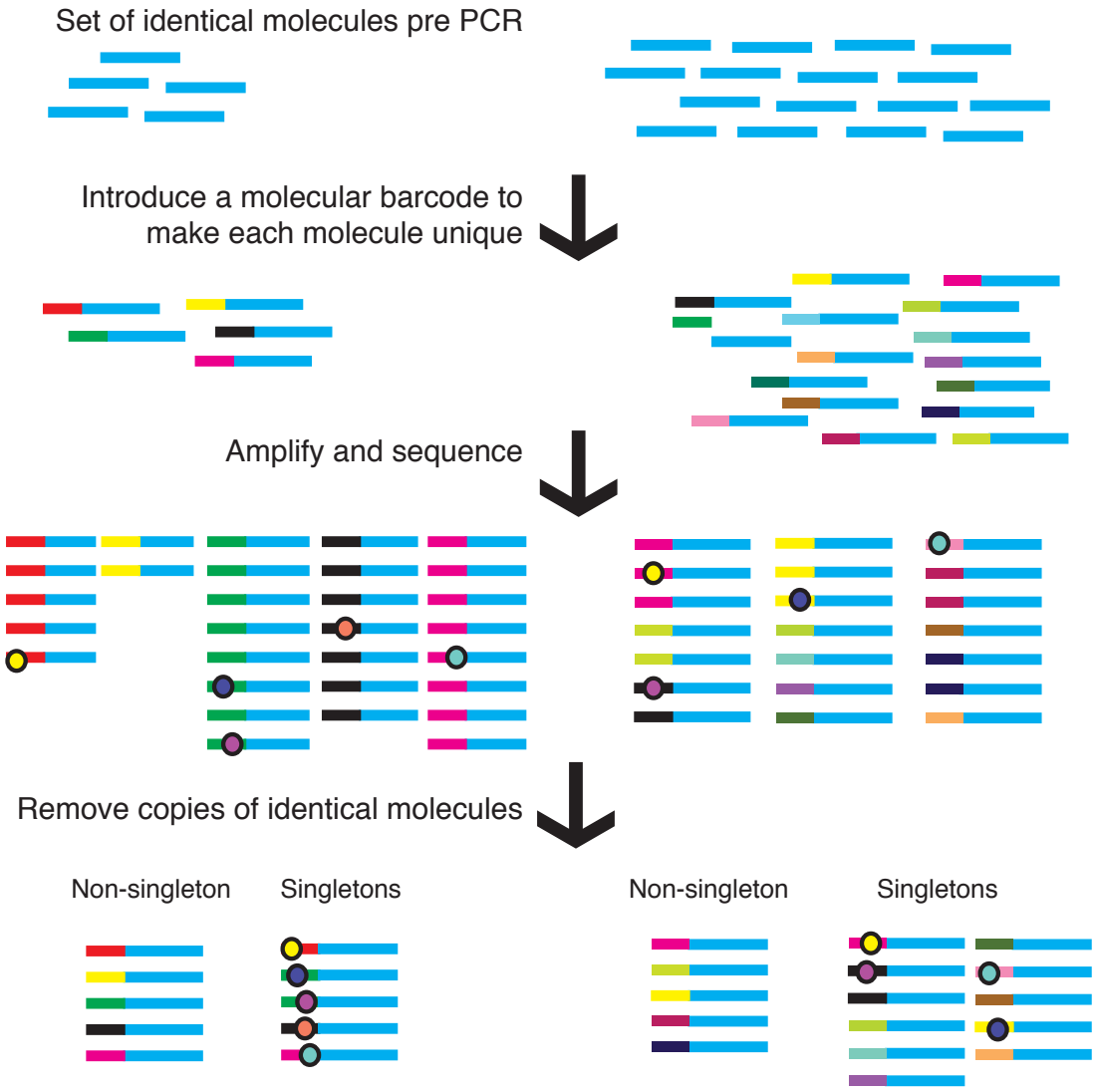

Figure 9: Singleton UMI sequences
On the left hand side few molecules are highly amplified and sequenced. Since each molecule is sequenced several times, most singletons are due to amplification or sequencing induced errors, either in the barcode or the template. However if the library isn't over sequenced, many non-mutant singleton molecules will also be detected.

2.3 APPLICATIONS OF MOLECULAR BARCODING

Apart from the experiments explained above the barcoding strategy has been used in many other applications. Some of these are discussed below to indicate the general applicability of the method.

2.3.1 Exact quantification of genomic copy number variation

Molecular barcoding of individual molecules have been applied to determining chromosomal copy number variation either with a fixed set of barcodes (49) or by endogenous barcodes and limiting dilution (34).

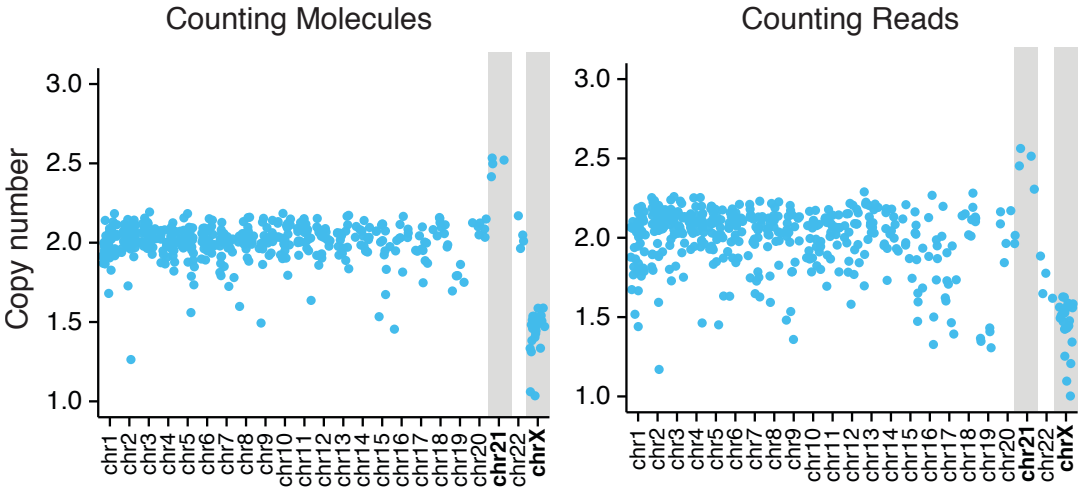

Figure 10: UMIs reduce noise in karyotyping
Copy numbers in sliding windows along the genome, in chromosome order. On the left, molecules (i.e. UMIs) were used to identify the karyotype. On the right, reads were used. Raw reads were mapped to the genome and UMI created based on the 5' mapping position. Reads and molecules were counted at 5 Mbp intervals and then centred around 2. Intervals around the centromeres and edges of chromosomes were removed. Half of the input DNA came from a boy with trisomy 21 and half from a healthy woman. Grey bars demarcate chromosome 21, which should have a copy number of 2.5 and chromosome X, which should have a copy number of 1.5. Adapted from Kivioja et al (2012).

Kivioja et al applied the endogenous method first demonstrated by Kinde et al to successfully determine the copy number status in a mixture of 50% DNA from a boy with trisomy 21 and 50% DNA from his mother. Here the whole genome was sequenced in contrast to earlier methods where specific amplicons or enrichments of parts of the genome were sequenced. Using endogenous barcodes the noise level was drastically reduced compared to just counting reads. It was also shown that deeper sequencing and counting reads didn't lower the noise and that using endogenous barcodes the coefficient of variation (CV, standard deviation / mean) was close to the theoretical maximum accuracy obtained by uniform random sampling. Data from this paper was used to create

figure 10, which shows that molecule counting reduces noise and is more suitable to identify the extra copy at chromosome 21.

2.3.2 Reducing noise in single-cell RNA sequencing

The usage of UMI to reduce noise in single cell RNA sequencing has become common practice for methods that capture either the 5' end like in single-cell tagged reverse transcription (STRT) (34, 58, 59) or the 3' end of molecules like in massively parallel RNA single-cell sequencing (MARS-seq) (60). Also Cell Expression by Linear amplification and Sequencing (CEL-seq) (61) has implemented the UMI in its latest protocol (62).

In STRT single cells are lysed and a polyT-primer is hybridized and allowed to extend with a reverse transcriptase that has template-switching capabilities. A template switching oligo is added with a 5' Illumina adapter sequence. Following amplification a second adapter is introduced by a transposon. The 3' side is cut with a restriction enzyme, allowing only the 5' side to amplify in a second round of amplification. After purification the library is ready for sequencing.

Zeisel et al used the STRT method to characterize the cellular composition of the mouse somatosensory cortex and hippocampal Cornu Ammonis area 1 (CA1) region (63). Regions of interest of the mouse brain were dissected, dissociated and loaded into the Fluidigm C1 instrument. The C1 instrument allows for single cell isolation, lysis and sequencing preparation with the STRT protocol. 3005 single cells were retained after quality control and used to identify 9 main classes of cells, and 47 distinct molecular subclasses by a novel biclustering method called backSPIN, that clusters genes and cells simultaneously. New specific markers for each class and subclass were identified extending the number of known markers that can be used to identify cells. It was also found that transcription factors formed a layered regulatory code suggesting their role in maintaining cell state.

Figure 11 shows an example where UMI is used to reduce technical variability in a single-cell RNA experiment. By adding an equal amount of molecules to each cell in the single cell experiment, technical variation can be estimated. This is commonly done using a standard set of molecules of varying length and concentration from the External RNA Controls Consortium (ERCC). Genes that have a higher variation than the spike-in molecules show biological variation on top of the technical variation, and are therefore of particular interest e.g. when identifying cell types.

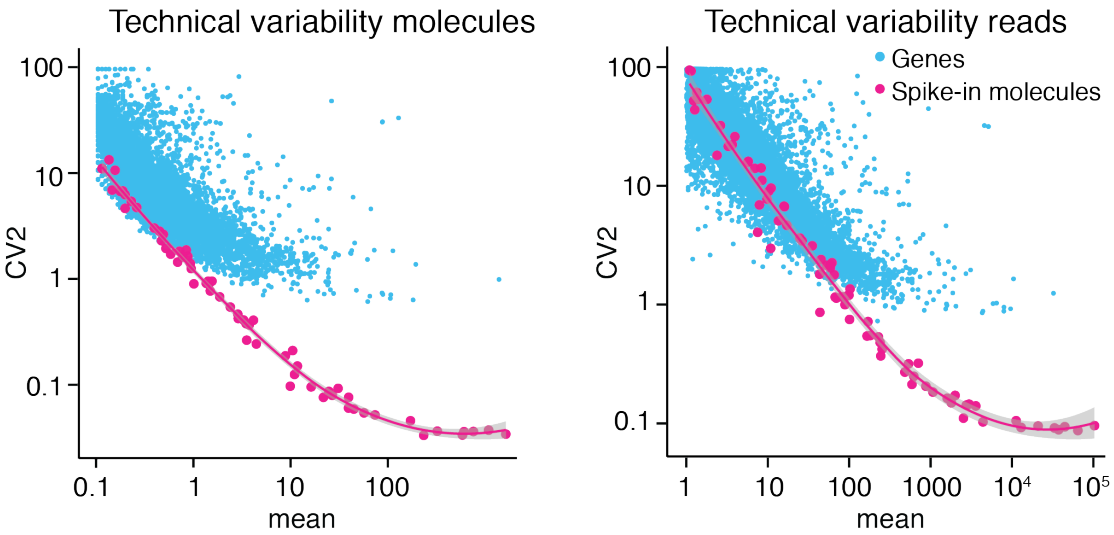

Figure 11: UMIs reduce noise in single-cell RNA-seq
Spike-in molecules can be used to model technical variability. If they are added in the same concentration to each cell before library preparation they can be used as a proxy for technical variation. Genes are shown in blue and spike-in molecules in pink. The pink line is a line fitted to the spike-in molecules. Data is taken from (63). On the left, expression was measured in molecule counts (i.e. UMIs). On the right, expression was measured in read counts. Using UMIs reduced technical noise, and enabled detection of true biological variability in a greater number of genes.

2.3.3 Assessment of efficiency in library preparation

Fu et al used molecular barcoding to assess the efficiency of standard Illumina library preparation for RNA sequencing and concluded that for every 1000 copies of a transcript in the starting sample only 1-6 copies, depending on the ERCC molecule used in the measurement, remained in the sequencing library (64). In the same study they also showed that it isn't enough to just use the start and end position of a randomly fragmented RNA molecule as an endogenous barcode since the fragmentation (or random priming or reverse transcription) isn't truly random. There are many more molecules with identical start and end positions than would be expected if the break point would assume a uniform distribution.

The molecular barcoding strategy has also been put into practice for example in RNA-proteins interactions using the UV cross-linking and immunoprecipitation (iCLIP) method (39), and in lineage tracing (65, 66).

2.4 AN ALTERNATIVE METHOD TO CORRECT FOR ERRORS

If input material is plenty an alternative to error correction with a barcoding strategy would be to prepare a library without amplification. Such a library would per definition not contain any amplification-induced errors, but it would still be prone to other errors created during the library preparation.

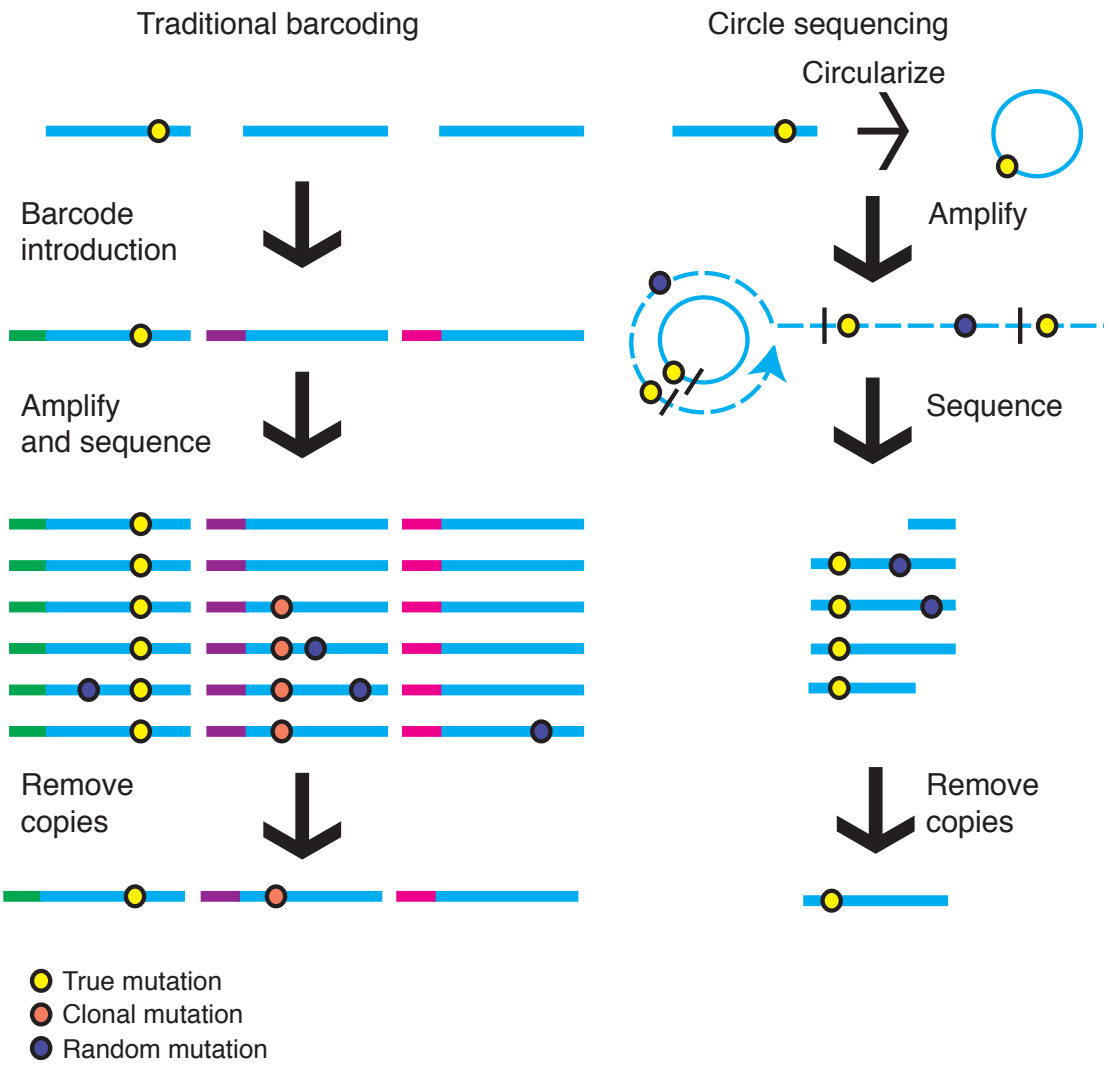

Figure 12: Circle sequencing to correct for errors
A major difference between traditional barcoding, shown to the left, and circle sequencing, shown to the right, is that traditional barcoding in combination with PCR amplification leads to both clonal and random errors, where circle sequencing only has random errors. Adapted from Lou et al (2013).

Lou et al proposed an alternative method, Circle Sequencing, to correct for library preparation and sequencing errors (67). Instead of making each molecule unique, the library was size-selected to contain only molecules around 1/3 of expected sequencing length. Molecules were then circularized and amplified with rolling circle amplification (RCA) before sequencing as shown in figure 12.

A major benefit using RCA is that artefactual mutations in the copy aren't clonally expanded, as is the case of mutations introduced by PCR. Instead mutations are randomly introduced and are therefore easy to correct for. In contrast to Duplex Sequencing, damaged DNA may still propagate errors. These errors are reduced in Circle Sequencing by treatment of DNA with uracil-DNA glycosylase and formamidopyrimidine-DNA glycosylase, which excise deaminated cytosine and 8-oxo-guanine bases, prior to library preparation. This treatment reduced the error rate by more than one order of magnitude and should be applicable in other molecular barcoding or library preparation protocols too.

Importantly yield, defined as the total number of high-quality consensus bases produced divided by the raw number of sequenced bases was substantially higher for Circle Sequencing compared to any other barcoding method. Standard UMI barcoding methods are sensitive to the amount of input material used: low amount of input material compared to sequencing depth results in redundant information, while a high amount of input material will give too few reads per barcode to build a consensus sequence. In contrast in Circle Sequencing the repeats are linked, and therefore reads per barcode will be the same regardless of the number of input molecules. It was noted however that for some applications a PCR amplification step might be necessary before Circle Sequencing, and in those cases Circle Sequencing will not be able to account for mutations acquired before circularization.

2.5 AN ALTERNATIVE METHOD TO CREATE UNIQUE SEQUENCES

So far two methods have been proposed to make identical molecules unique: endogenous and exogenous barcoding. In these two methods the barcode is placed at the end of the molecule allowing only sequences to be identified within one read length from the end. This creates a problem e.g. in mRNA isoform determination using short read assembly. It is possible to first fragment the mRNA and then add the UMI, however this would resort to a probabilistic model of isoform usage since the information of long-range exon connectivity is lost. Recently a paper was published that proposed a way to make molecules unique by exploiting incomplete bisulphite conversion of cDNA, leading to restricted random mutation before amplification (68). The idea is illustrated in figure 13.

The authors created a model where they varied the number of C residues, coverage per template, and conversion rate of C to T for reads with length of 2*100 bp. They concluded that it was enough with 30C residues and a flip rate 0.35 to make more than thousands of identical molecules unique. This assembly by mutagenesis would be beneficial not only in isoform determination but also for discrimination of haplotypes and de novo haplotype assembly.

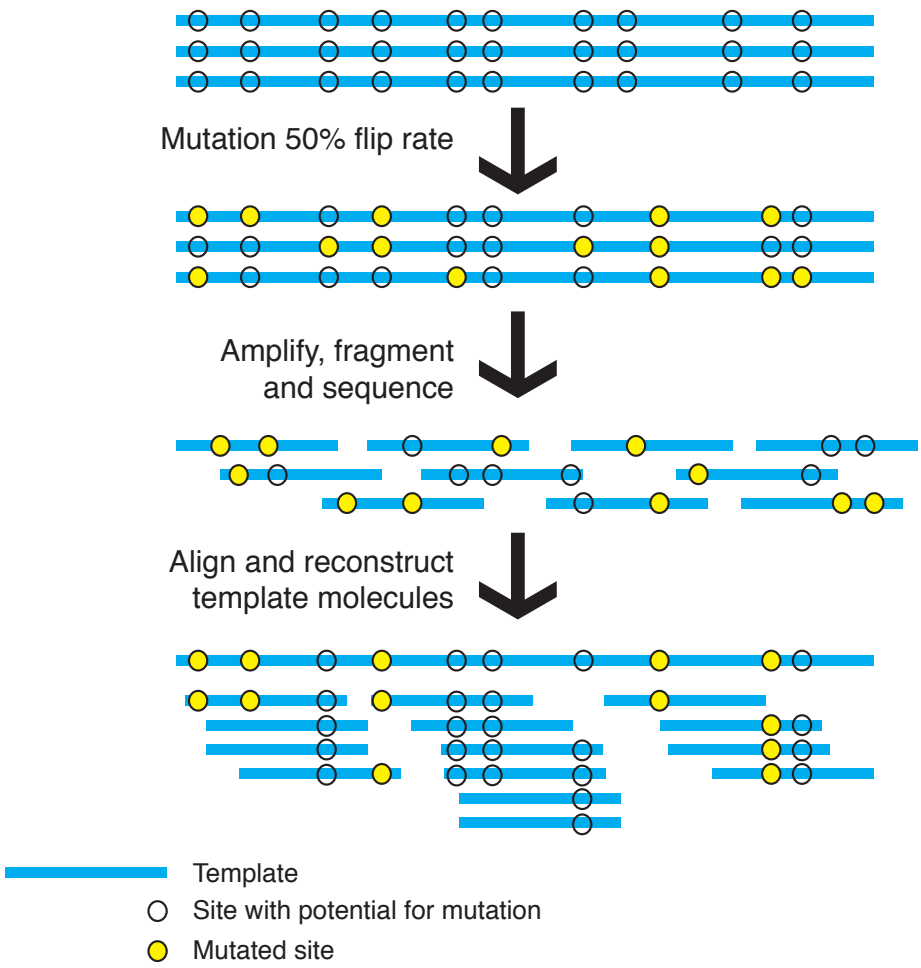

Figure 13: Template mutagenesis to make molecules unique
Identical templates can be made unique by exposing them to a mutation-inducing agent that mutates specific sites with a certain flip rate. Templates can then be amplified, fragmented and sequenced. After sequencing it is possible to both quantify the number of input template molecules and to stich the short reads together to re-create the long template molecule.

2.6 CONCLUSIONS PART 2

Molecular barcoding of individual molecules has been shown to increase quantification and reproducibility, reduce noise and lower the substitution, insertions and deletion errors produced by sequencing and amplification. Molecular barcoding has been used in many applications including sequencing of gut bacteria, lineage tracing and estimation of polymerase fidelity and in a number of different sequencing settings, like amplicon sequencing, DNA sequencing from a genomic region, and in single-cell RNA sequencing.

The length of the barcode differs substantially between different experiments and depends on the expected number of identical template molecules. Grün et al found that using a molecular barcode as small as 4 nt reduced technical noise in single-cell RNA sequencing for almost all genes with about 50% on average, compared to not using the barcode (69).

There are a number of pitfalls using molecular barcodes. First, errors in the UMI can easily cause an artificially high estimation of number of molecules. The longer the UMI and the more rounds of amplification used the higher the probability of UMI artefacts. Three approaches have been taken to solve this problem. UMI molecules with single or few reads can be removed. Amplification errors are clonal but most of them arise in the late rounds of amplification and are usually not represented by more than one or a couple of sequencing reads, although this depends of the sequencing depth. In comparison sequencing errors are random and are therefore often only represented by on one read. Both can therefore often be removed by filtering away UMI's with low read count. Another, and complementary, approach is to collapse UMIs with a Hamming distance of 1 or 2 depending on the length of the UMI. A prerequisite of this approach is that there are enough random barcodes in relation to the number of template molecules so that it is very unlikely that two UMIs have a Hamming distance of 2 or less. The third approach is to use UMI with a defined barcode set with a known Hamming distance. This is feasible for paired end sequencing as shown by Shiroguchi et al.

The second pitfall has to do with collision events. It is important that the number of barcodes is in large excess of the number of identical template molecules. Generally, it is good if the number of template molecules are fewer than the square root of the number of barcodes (40). Even with fewer barcodes it is possible to make an estimation of the number of input templates based on the read distribution of those barcodes, and generally exact quantification becomes less crucial for abundant templates.

The third pitfall has to do with redundant sequencing. UMIs should preferentially be sequenced to saturation for exact quantification and error correction. In practice this means that each molecule should be sequenced at least once to be sure that all molecules have been detected, and preferably more times to allow for error correction. This requirement reduces the yield of barcoding strategies compared with standard sequencing methods.

It is also important to note that although standard UMI barcoding will make it possible to measure molecules in an absolute scale, there are often other biases in the library. E.g. in single-cell RNA sequencing UMI label cDNA, which means that RNA templates not converted to cDNA will not be counted. Also low

abundance molecules will be stochastically amplified and sequenced leading to a bias in representation, and early PCR errors, e.g. from damaged template molecules, will not be corrected unless Duplex Sequencing is applied.

Since the introduction of single molecule barcoding in 2003, a number of improvements have been made. Casbon et al showed that UMI could be used for error correction. Kinde et al suggested that endogenous UMI also could be used. Shugay et al identified common mutations in late rounds of PCR and showed that these mutational hotspots could be used to correct for errors occurring in the first rounds of PCR. Schmitt et al showed that using information from both strands of DNA dramatically lowered amplification errors since it could distinguish the errors created in the first round of PCR from true mutations. Shiroguchi et al introduced a defined set of barcodes that were different from each other on several positions to severely reduce the number of false positives due to mutations in the UMI. Faith et al increased reproducibility by shifting the bottleneck to the amount of primers used instead of on the amount of input DNA. Lundberg et al proposed that some singleton UMI molecules could actually be useful: singletons in libraries with few reads per UMI are usually more correct and informative than singletons in libraries with many reads per UMI. A competing method for error correction has been developed as well as a new method for creating unique templates.

Currently there is a trade-off between the quality of the error correction and yield. The choice of molecular barcoding strategy will depend on the application at hand. Not all methods can be used for all applications: e.g. duplex sequencing is not feasible for RNA-sequencing since RNA is single stranded and Circle Sequencing is not applicable to low input material. Features of the different methods are summarized in Table 1.

In conclusion molecular barcoding of individual molecules has been shown to be efficient at reducing amplification-induced quantification bias and to correct for errors produced by library preparation and sequencing in many areas of molecular biology and it is likely to be even more applied in the future.

Table 1. Properties of molecular barcoding strategies

	Normal PCR based library prep	Safe-SeqS	MIGEC	Duplex Sequencing	Circular Sequencing	Amplification-free library prep
Affected by PCR errors	Yes	Few	Few	Very few	No	No
Affected by sequencing errors	Yes	No	No	No	No	Yes
Affected by DNA damaged in one strand	Yes	Yes	Yes	No	Yes*	Yes
Yield	High	Low	Low	Very Low	High	Very High
Input material required	Low	Low	Low	Low	High	Medium
Accurate quantification	No	Yes	Yes	Yes	Yes	Yes
Introduction of uninformative molecules	Yes	Yes	Yes	Yes	No**	No
Even representation of genomic regions	No	No	No	No	Yes	Yes

* Reduced by treatment with uracil-DNA glycosylase and formamidopyrimidine-DNA glycosylase
** Still redundant information is contained within each read

3 RESULTS

3.1 PAPER I: COUNTING ABSOLUTE NUMBERS OF MOLECULES USING UNIQUE MOLECULAR IDENTIFIERS

Paper I introduces the concept of Unique Molecular Identifiers (UMI) in RNA sequencing and in karyotyping of genomic DNA. The paper suggests two different methods of making molecules unique before amplification, by either degenerate barcoding or random fragmentation and dilution so each molecule receives a unique starting position in the genome with a high probability.

Paper I showed that using UMI in RNA sequencing substantially reduces noise. There was a marked improvement in correlation between genes sequenced after 15 cycles of amplification compared to 25 cycles of amplification when counting molecules instead of reads. This indicated that the additional 10 cycles of amplification didn't skew the representation of molecules, however the difference in copy number of molecules changed and created noise when counting reads.

Counting molecules instead of reads also showed a drastic reduction of noise when used in non-invasive prenatal testing (NIPT) of fetal karyotype, measured as coefficient of variation (CV, standard deviation / mean) between genomic regions. It was also shown that increased sequencing depth could not reduce noise when counting reads.

3.2 PAPER II: AMPLIFICATION-FREE SEQUENCING OF CELL-FREE DNA FOR NON-INVASIVE PRENATAL TESTING OF FETAL CHROMOSOMAL ABERRATIONS

Reduction of noise in sequencing data is of particular importance in the clinical setting where low accuracy can lead to an erroneous decision by a physician, which in turn can have fundamental consequences for the patient.

When cells die their DNA is fragmented and enters the blood stream. It has been known for a long time that this so called cell-free DNA (cfDNA) exists, and it has been tested for monitoring cancer progression. In 1997 Dennis Lo discovered that cfDNA from the foetus can be found in the maternal blood stream (70). This discovery led to some immediate clinical applications like tests for sex and Rhesus factor. With the advent of NGS, fetal cfDNA was also used for NIPT of fetal karyotype (71). This is complicated by the fact that fetal cfDNA is mixed in a much larger pool of maternal cfDNA. The accuracy of NIPT methods were

dependent of the chromosome examined and most methods performed worse on chromosomes with aberrant GC content.

It was hypothesized that molecular barcoding would increase accuracy of NIPT leading to more correct clinical decisions. However clinical reality has many parameters to take into account and it was soon discovered that the low yield of the UMI method would make it too costly to implement. Therefore an amplification-free library preparation protocol was developed that successfully produced libraries from the small amount of cfDNA extractable from plasma samples. If amplification is removed from the library preparation a large part of the bias associated with it is removed. An added benefit of amplification-free protocols is that each read is informative, that is no non-informative copies of a molecules are sequenced, which in turn allows for fewer reads to be sequenced.

Paper II showed that the amplification-free library preparation method could be used to correctly identify the karyotype of 27 foetuses, of which 15 had one or more aberrant karyotypes, using cfDNA from maternal plasma. It also showed that both the amplification-free library preparation and the UMI method substantially lowered bias in karyotyping when a single sample was mapped to the genome. However when a sample was normalized to a control there was no clear benefit neither for the amplification-free nor the UMI method. This phenomenon is likely due to that the bias introduced by PCR is highly conserved between samples, and when a sample produced with a standard amplification protocol is compared with another sample the bias can be identified and accounted for.

3.3 PAPER III: ALTERNATIVE PROMOTERS ARE CO-REGULATED IN SINGLE CELLS IN THE MOUSE BRAIN

An organ consists of a multitude of single cells that interact in complex ways. A reductionist view of studying organ function would be to study the individual parts of the organ, i.e. the single cells making up the organ. Until recently this has been very difficult to do, apart from studying the morphology or individual mRNA transcripts or proteins, and describing a cell solely on morphology is clearly inadequate. Recent advances in single-cell RNA-sequencing is about to change this, and there is an on-going debate in the field on how to define a cell type. Studying biology at the single-cell level increases the resolution of the process studied, compared to studying a mixture of cells at the same time. E.g. a gene moderately expressed at the bulk level could either be moderately expressed in all cells or be highly expressed in some cells and not expressed in others.

The definition of a promoter is a "DNA sequence(s) that define where transcription of a gene by RNA polymerase begins" (72). Promoters are typically located directly upstream of the TSS, and are bound by the RNA polymerase and with it associated factors to initiate transcription. Initially it was thought that a gene has only one promoter, but it has now been shown that a gene can have several promoters (73). Since the advent of NGS promoter usage has been studied in detail, however it has not been done in single cells.

Using STRT in combination with UMIs, Paper III showed that if a gene expresses transcripts from two promoters in the bulk population it usually expresses transcripts from both promoters also in single cells. Interestingly these two promoters are generally expressed in a conserved ratio across cells of a specific cell type, in contrast with gene expression level that varies considerably between cells. Typically the ratio between the major promoter and the minor promoter are also conserved across cell types, but here specific genes can change the ratio of expression, and in some cases the minor promoter is also higher expressed than the major. When comparing two different neuronal cell types few genes significantly change the ratio of promoter expression, however when comparing a neuronal and a non-neuronal cell type in the mouse brain, the change of expression is more common, indicating that genes governing the neuronal phenotype also influence promoter preference. A major conclusion from this paper is that promoter expression in a cell type is generally governed by a common factor that influences both the major and minor promoter, although with different affinity. This conclusion is somewhat supported by the discovery that neighbouring promoters often interact with each other (74) and that active promoters tend to influence the expression of neighbouring genes (75).

3.4 PAPER IV: SINGLE-CELL mRNA ISOFORM DIVERSITY IN THE MOUSE BRAIN

Alternative isoform usage is known to be an important feature of our genome and allows it to create a highly diverse set of proteins from a relatively small number of genes. Two transcripts from the same gene can differ from each other in many ways. Their start or end position could differ, potentially leading to different degradation kinetics. Their exon composition could vary, or transcripts could retain introns during splicing, leading to severely altered proteins. Finally they could vary in exon end or start positions.

In theory each gene could be represented by many different isoforms, and there are certain indications that they indeed are. A recent study by the Encode consortium found that the number of expressed isoforms in a cell line tended to follow the number of annotated isoforms up to 10-12 isoforms expressed per

gene (76). Another study by the FANTOM consortium found that on average human genes have four robust CAGE peaks within 500 bp from the annotated 5' end of the gene (73). Both these studies have been done on bulk material and the amount of isoform diversity at the single cell level is unknown.

Paper IV takes advantage of the PacBio long read sequencing platform and the increased accuracy achieved with UMI molecule counting to examine full length mRNA for the whole transcriptome at the single cell level. The major finding in paper IV is that a large part of all mRNA molecules in a single cell constitute separate isoforms, even after applying a conservative definition of what constitutes an isoform. Also relatively few isoforms are common between cells. Another finding was that exon junctions in coding regions show less isoform diversity than exon junction in non-coding regions. In conclusion genes express a surprisingly high number of isoforms also in single cells, which indicates that the transcriptional machinery can afford to be a bit inaccurate, especially outside of the coding region.

4 PERSPECTIVES

Molecular biology is in an intensive phase of making new discoveries. Two events stand out as being especially important in providing a framework for new innovations: The sequencing of the human genome and the advent of next generation sequencing. The development of methods to get information on the genome, transcriptome and even proteome at the single cell level has shown that single cells are more heterogeneous than previously thought, which will have an impact both on basic research and medicine.

The early days of molecular biology were preoccupied with figuring out how the flow of information went from DNA to RNA to protein. Today a lot of effort is spent on understanding the regulation of those steps, how transcription factors and other epigenetic elements combine to turn on and off genes.

Although it is always risky to predict the future some general trends can be discerned that will likely influence molecular biology in the coming years. Many efforts are ongoing to make sequencing cheaper, easier, quicker and with longer reads. Today Illumina's short read technology is dominating the sequencing market, and it is used in a number of applications from targeted resequencing and diagnostics, to de novo genome assembly. Other technologies are now maturing and it seems likely that some of these technologies will take over certain sectors of the market, like *de novo* genome sequencing and in diagnostic applications where time or convenience is a limiting factor.

Another trend is that single cell sequencing technologies will grow in popularity. During the last five years single-cell RNA sequencing has transformed from being performed by a few laboratories, and limited to sequencing a few cells, to becoming a routine procedure offered by core facilities. A number of commercial alternatives now exist that can prepare single-RNA sequencing libraries of up to 48,000 cells at a time (77). Single-cell experiments are also likely to measure a growing number of parameters per cell, such as combining transcriptomics and proteomics, or transcriptomics and DNA sequencing (78, 79). Another trend is that single cell sequencing is done *in situ*, which has already been demonstrated by several groups (80, 81). So far most single-cell RNA-sequencing experiments have been concerned with characterizing the normal state of cells, but more and more studies are done on perturbed states, like in disease or after exposure to a drug. I believe the frontier of single cell experiments will soon move on to examination of single cells in their natural environment, and cell-cell interactions, since expression has been shown to be strongly influenced by other cells and the local niche.

Molecular barcodes are already commonly used in single-cell RNA sequencing experiments. They have been successfully used in targeted resequencing and other fields of molecular biology to increase accuracy of sequencing, and will surely get more generally accepted by the sequencing community.

Finally, I believe the coming years will see many more examples of sequencing being used in the clinic. Currently monogenic diseases are being screened and chromosomal aberrations are investigated in cancer and prenatal diagnostics. Coming genetic tests may include testing total viral load in blood and gut bacteria composition. Liquid biopsies of the blood is also coming of age, where circulating tumour cells, tumour-related cell-free DNA, or exosomes can be studied, allowing the doctor to on a daily basis see the genetic part of disease progression. With refined diagnostic tools personalized medicine will be transformed from anecdotal to standard medical practice.

5 ACKNOWLEDGEMENTS

First and foremost, thank you **Sten**! It's a rare gift to meet someone who is able to transform you into a wiser person, and an even more rare gift to work together with that person on a daily basis for five years.

Thank you to my supervisors **Magnus Nordenskjöld** and **Erik Iwarsson** for teaching me the intricate details of prenatal diagnostics. And to **Kalle Malmberg,** even if our project never materialized I had a good time talking to you about natural killer cell biology. Thank you **Jan-Inge Henter** for inspiring meetings, which have given me fruitful thoughts and definitely influenced some of my decisions.

Thank you **Itai Yanai** for crossing the Atlantic to be my opponent and making this dissertation an unforgettable event for me! And to **Mats Nilsson**, **Carsten Daub** and **Per Uhlén** for participating in the board of examination, and to **Ulrika Marklund** for being the chairman of the disputation process. It's an honour!

A warm thank you to past and current members of the Linnarsson group.
Saiful for patiently introducing me to laboratory work, **Peter** for doing the same for programming. Thank you **Anno Jo** for your passionate opinions on all matters and **Anna Ju** for never mixing up the primers. Thank you **Una** for all heated debates and **Pawel** for your cheerful mood. **Amit** for your deep commitment to science, your arrival really influenced the lab to the better, and **Gioele** for your passion for research and great ideas. Thank you **Simone** for your happy energy, it's been great to have an artisan in the group, and **Lars** for your patient optimism. Thank you **Hannah** for being the social one, it has been very pleasant to be your desk neighbour. Thank you **Abeer** and **Rickard** and the **Core Facility** for all your hard work.

A warm thank you to past and current members of MolNeuro. Thank you **Patrik**, **Ernest**, **Per**, **Gonçalo, Jens,** and **Ulrika**, you have managed to create a great research environment. I have enjoyed my five years here very much and even if it's not always noticed, a big part has to do with the structures built by you. Thank you **Johnny** and **Alessandra** for your friendly attitude and all your support. Thank you **Satish** for all your inspiration and ideas, and you **Enrique** for making MolNeuro a bit more cheerful and you **Daniel** my companion since the stem cell course. Thank you **Ivar** for cheering up our room and letting me in on some of your world-changing ideas. And thank you **Lili** for trying to make me speak your native language, and you **Maryam** for all good laughs, and **Spyros** for being so upliftingly positive. Thank you **Sueli**, **Ana**, **Nina**, **Shanzheng**, **Songbai**, **Alca**, **Carlos**, **Carmen**, **Lottie**, **Connla**, **Sam**, **Daohua**, **Mitya**, **Puneet, Dagmara**, **Fatima**, **Erik**, **Staffan** and all others for making this time so memorable.

Thank you **Ellika** and **Leo** for trying to shoot the moon with me, I learnt a lot from you! And thank you **Patrik Blomquist** for your support using both carrots and sticks, usually wrapped in cotton.

Thank you **Leo** for agreeing on almost nothing and still being a good friend. Amazing such a thing is still possible. And to you **Niko** for your bubbling enthusiasm about science and life, and to both of you for excellent lunch discussions covering both the profound and mundane. Thank you **Petter** for great friendship and for not just being good. And thank you **Viking** for being so curious.

To my parents, **Maria** and **Christer**, and to **Lennart**, thank you very much! Without your support in critical moments my work would have been much more difficult. And I would like to thank my wonderful wife **Dongjiao** who travelled half the world to live with someone she didn't know would end up a scientist, often absent minded, lost in thoughts. You have supported me through long hours and early mornings, it has been great to have you by my side! And finally a thank to my son **Numa,** who is as stubborn in his cravings as uncompromising in his love.

6 REFERENCES

1. Watson HDC, F.HC. . Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature. 1953;171:737-8.
2. Crick FHC. On protein synthesis. Symp Soc Exp Biol. 1958;12:138–63.
3. Crick FHC. Central Dogma of Molecular Biology. Nature. 1970.
4. Choudhuri S. Some major landmarks in the path from nuclein to human genome (1). Toxicology mechanisms and methods. 2006;16(2-3):137-59. PubMed PMID: 20021005.
5. Loening UE. The Fractionation of High-Molecular-Weight Ribonucleic Acid by Polyacrylamide-Gel Electrophoresis. Biochem J. 1967;102:251-7.
6. Southern EM. Detection of Specific Sequences Among DNA Fragments Separated by Gel Electrophoresis. J Mol Biol. 1975;98:503-17.
7. Alwine JC, Kemp DJ, Stark GR. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc Natl Acad Sci. 1977;74:5350-4.
8. Smith HO, Wilcox KW. A restriction enzyme from Hemophilus influenzae - I. Purification and general properties. J Mol Biol. 1970;51:397-1.
9. Weiss B, Richardson CC. Enzymatic breakage and joining of deoxyribonucleic acid, I. Repair of single-strand breaks in DNA by an enzyme system from Escherichia coli infected with T4 bacteriophage. Proc Natl Acad Sci. 1967;57(1021-8).
10. Temin HM, Mizutani S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature. 1970;226.
11. Baltimore D. Viral RNA-dependent DNA Polymerase. Nature. 1970;226.
12. Jackson D, Symons R, Berg P. Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40- Circular SV40 DNA Molecules Containing Lambda Phage Genes and the Galactose Operon of Escherichia coli. Proc Nat Acad Sci. 1972;69(10):2904-9.
13. Cohen S, Chang A, Boyer H, Helling R. Construction of biologically functional bacterial plasmids in vitro. Proc Nat Acad Sci. 1973;70(11):3240-4.
14. Jaenisch R, Mintz B. Simian Virus 40 DNA Sequences in DNA of Healthy Adult Mice Derived from Preimplantation Blastocysts Injected with Viral DNA. Proc Nat Acad Sci. 1974;71(4):1250-4.
15. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci. 1977;74:560-4.
16. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94:441-8.
17. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. 1977;74:5463-7.
18. Saiki RK, Scharf S, Faloona F, K.B. M, Horn GT, Erlich HA, et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science. 1985;230:1350-4.
19. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921. PubMed PMID: 11237011.
20. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001 Feb 16;291(5507):1304-51. PubMed PMID: 11181995.

21. Fenno L, Yizhar O, Deisseroth K. The development and application of optogenetics. Annual review of neuroscience. 2011;34:389-412. PubMed PMID: 21692661.
22. Boyden ES, Zhang F, Bamberg E, Nagel G, Deisseroth K. Millisecond-timescale, genetically targeted optical control of neural activity. Nature neuroscience. 2005 Sep;8(9):1263-8. PubMed PMID: 16116447.
23. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012 Aug 17;337(6096):816-21. PubMed PMID: 22745249.
24. Shendure J, Ji H. Next-generation DNA sequencing. Nature biotechnology. 2008 Oct;26(10):1135-45. PubMed PMID: 18846087.
25. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008 Nov 6;456(7218):53-9. PubMed PMID: 18987734. Pubmed Central PMCID: 2581791.
26. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science. 2009;323:133-8. PubMed PMID: 19023044.
27. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. PubMed PMID: 22955616. Pubmed Central PMCID: 3439153.
28. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome biology and evolution. 2013;5(3):578-90. PubMed PMID: 23431001. Pubmed Central PMCID: 3622293.
29. Invitrogen. http://tools.thermofisher.com/content/sfs/brochures/711-021834 AccuPrime Brochu.pdf. (Retrieved 2016-01-12).
30. Kapabiosystems. https://http://www.kapabiosystems.com/product-applications/products/pcr-2/kapa-hifi-pcr-kits/, (Retrieved 2016-01-12).
31. Dabney J, Meyer M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. BioTechniques. 2012 Feb;52(2):87-94. PubMed PMID: 22313406.
32. Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic acids research. 2015 Dec 2;43(21):e143. PubMed PMID: 26187991. Pubmed Central PMCID: 4666380.
33. Vogelstein B, Kinzler KW. Digital PCR. Proc Natl Acad Sci. 1999;96:9236-41.
34. Kivioja T, Vaharautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. 2012 Jan;9(1):72-4. PubMed PMID: 22101854.
35. Karlsson K, Sahlin E, Iwarsson E, Westgren M, Nordenskjold M, Linnarsson S. Amplification-free sequencing of cell-free DNA for prenatal non-invasive diagnosis of chromosomal aberrations. Genomics. 2015 Mar;105(3):150-8. PubMed PMID: 25543032.
36. Hug H, Schuler R. Measurement of the Number of Molecules of a Single mRNA Species in a Complex mRNA Preparation. Journal of Theoretical Biology. 2003;221(4):615-24.

37. Miner BE, Stoger RJ, Burden AF, Laird CD, Hansen RS. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. Nucleic acids research. 2004;32(17):e135. PubMed PMID: 15459281. Pubmed Central PMCID: 521679.
38. McCloskey ML, Stoger R, Hansen RS, Laird CD. Encoding PCR products with batch-stamps and barcodes. Biochemical genetics. 2007 Dec;45(11-12):761-7. PubMed PMID: 17955361.
39. Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nature structural & molecular biology. 2010 Jul;17(7):909-15. PubMed PMID: 20601959. Pubmed Central PMCID: 3000544.
40. Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP. A method for counting PCR template molecules with application to next-generation sequencing. Nucleic acids research. 2011 Jul;39(12):e81. PubMed PMID: 21490082. Pubmed Central PMCID: 3130290.
41. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2011 Jun 7;108(23):9530-5. PubMed PMID: 21586637. Pubmed Central PMCID: 3111315.
42. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proceedings of the National Academy of Sciences of the United States of America. 2011 Dec 13;108(50):20166-71. PubMed PMID: 22135472. Pubmed Central PMCID: 3250168.
43. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2012 Sep 4;109(36):14508-13. PubMed PMID: 22853953. Pubmed Central PMCID: 3437896.
44. Hensel MS, J.E; Gleeson,D.;Jones,M.D.;Dalton,E.;Holden,D.W. Simultaneous Identification of Bacterial Virulence Genes by Negative Selection. 1995.
45. Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. Nat Genet. 1996 Dec;14(4):450-6. PubMed PMID: 8944025.
46. Qiu F, Guo L, Wen TJ, Liu F, Ashlock DA, Schnable PS. DNA sequence-based "bar codes" for tracking the origins of expressed sequence tags from a maize cDNA library constructed using multiple mRNA sources. Plant physiology. 2003 Oct;133(2):475-81. PubMed PMID: 14555776. Pubmed Central PMCID: 523874.
47. Cox JPL. Bar coding objects with DNA. The Analyst. 2001;126(5):545-7.
48. Cook LJ, Cox JPL. Methylated DNA labels for marking objects. 2002.
49. Fu GK, Hu J, Wang PH, Fodor SP. Counting individual DNA molecules by the stochastic attachment of diverse labels. Proceedings of the National Academy of Sciences of the United States of America. 2011 May 31;108(22):9026-31. PubMed PMID: 21562209. Pubmed Central PMCID: 3107322.
50. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. Nat Methods. 2014 Jun;11(6):653-5. PubMed PMID: 24793455.
51. Egorov ES, Merzlyak EM, Shelenkov AA, Britanova OV, Sharonov GV, Staroverov DB, et al. Quantitative profiling of immune repertoires for minor

lymphocyte counts using unique molecular identifiers. Journal of immunology. 2015 Jun 15;194(12):6155-63. PubMed PMID: 25957172.
52. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. Nature protocols. 2014 Nov;9(11):2586-606. PubMed PMID: 25299156. Pubmed Central PMCID: 4271547.
53. Kennedy SR, Salk JJ, Schmitt MW, Loeb LA. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS genetics. 2013;9(9):e1003794. PubMed PMID: 24086148. Pubmed Central PMCID: 3784509.
54. Schmitt MW, Loeb LA, Salk JJ. The influence of subclonal resistance mutations on targeted cancer therapy. Nature reviews Clinical oncology. 2015 Oct 20. PubMed PMID: 26483300.
55. Shiroguchi K, Jia TZ, Sims PA, Xie XS. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proceedings of the National Academy of Sciences of the United States of America. 2012 Jan 24;109(4):1347-52. PubMed PMID: 22232676. Pubmed Central PMCID: 3268301.
56. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, et al. The long-term stability of the human gut microbiota. Science. 2013 Jul 5;341(6141):1237439. PubMed PMID: 23828941. Pubmed Central PMCID: 3791589.
57. Lundberg DS, Yourstone S, Mieczkowski P, Jones CD, Dangl JL. Practical innovations for high-throughput amplicon sequencing. Nat Methods. 2013 Oct;10(10):999-1002. PubMed PMID: 23995388.
58. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014 Feb;11(2):163-6. PubMed PMID: 24363023.
59. Islam S, Kjallquist U, Moliner A, Zajac P, Fan JB, Lonnerberg P, et al. Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. Nature protocols. 2012 May;7(5):813-28. PubMed PMID: 22481528.
60. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014 Feb 14;343(6172):776-9. PubMed PMID: 24531970. Pubmed Central PMCID: 4412462.
61. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell reports. 2012 Sep 27;2(3):666-73. PubMed PMID: 22939981.
62. Hashimshony T, Yanai I. http://www.dropbox.com/s/iwl30ss0nqmtddf/ CEL-Seq protocol_July_2014.docx. (Retrieved 2016-01-05).
63. Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015 Mar 6;347(6226):1138-42. PubMed PMID: 25700174.
64. Fu GK, Xu W, Wilhelmy J, Mindrinos MN, Davis RW, Xiao W, et al. Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. Proceedings of the National Academy of Sciences of the United States of America. 2014 Feb 4;111(5):1891-6. PubMed PMID: 24449890. Pubmed Central PMCID: 3918775.

65. Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G. Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature. 2015 Mar 12;519(7542):181-6. PubMed PMID: 25731169. Pubmed Central PMCID: 4426284.
66. Blundell JR, Levy SF. Beyond genome sequencing: lineage tracking with barcodes to study the dynamics of evolution, infection, and cancer. Genomics. 2014 Dec;104(6 Pt A):417-30. PubMed PMID: 25260907.
67. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, et al. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2013 Dec 3;110(49):19872-7. PubMed PMID: 24243955. Pubmed Central PMCID: 3856802.
68. Levy D, Wigler M. Facilitated sequence counting and assembly by template mutagenesis. Proceedings of the National Academy of Sciences of the United States of America. 2014 Oct 28;111(43):E4632-7. PubMed PMID: 25313059. Pubmed Central PMCID: 4217440.
69. Grun D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014 Jun;11(6):637-40. PubMed PMID: 24747814.
70. Lo YMD, Corbetta N, Chamberlain PF, Rai V, Sargent IL, Redman CWG, et al. Presence of fetal DNA in maternal plasma and serum. The Lancet. 1997;350(9076):485-7.
71. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proceedings of the National Academy of Sciences of the United States of America. 2008 Oct 21;105(42):16266-71. PubMed PMID: 18838674. Pubmed Central PMCID: 2562413.
72. Nature. http://www.nature.com/scitable/definition/promoter-259. Retrieved (2016-01-19).
73. Consortium F, the RP, Clst, Forrest AR, Kawaji H, Rehli M, et al. A promoter-level mammalian expression atlas. Nature. 2014 Mar 27;507(7493):462-70. PubMed PMID: 24670764. Pubmed Central PMCID: 4529748.
74. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012 Jan 20;148(1-2):84-98. PubMed PMID: 22265404. Pubmed Central PMCID: 3339270.
75. Rajagopal N, Srinivasan S, Kooshesh K, Guo Y, Edwards MD, Banerjee B, et al. High-throughput mapping of regulatory DNA. Nature biotechnology. 2016 Jan 25. PubMed PMID: 26807528.
76. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012 Sep 6;489(7414):101-8. PubMed PMID: 22955620. Pubmed Central PMCID: 3684276.
77. 10XGenomics. http://www.bio-itworld.com/2016/2/11/10x-genomics-reveals-upgraded-platform-new-features-single-cell-rna-sequencing.html. (Retrieved 2016-02-16).
78. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015 Jun;12(6):519-22. PubMed PMID: 25915121.

79. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. Nature biotechnology. 2015 Mar;33(3):285-9. PubMed PMID: 25599178. Pubmed Central PMCID: 4374170.
80. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015 Apr 24;348(6233):aaa6090. PubMed PMID: 25858977. Pubmed Central PMCID: 4662681.
81. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, et al. Highly multiplexed subcellular RNA sequencing in situ. Science. 2014 Mar 21;343(6177):1360-3. PubMed PMID: 24578530. Pubmed Central PMCID: 4140943.